MARYLAND

COLLEGE PARK CAMPUS

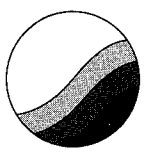
A GENERALIZED FINITE ELEMENT METHOD FOR SOLVING THE HELMHOLTZ EQUATION IN TWO DIMENSIONS WITH MINIMAL POLLUTION

by

Ivo M. Babuška
Frank Ihlenburg
Ellen T. Paik
and
Stefan A. Sauter

Technical Note BN-1179

September 1994



INSTITUTE FOR PHYSICAL SCIENCE AND TECHNOLOGY

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Note BN-1179

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): A Generalized Finite Element Method for Solving the Helmholtz Equation in Two Dimensions with Minimal Pollution

5. TYPE OF REPORT & PERIOD COVERED: Final Life of Contract

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): Ivo M. Babuska[1] - Frank Ihlenburg[2] - Ellen T. Paik - Stefan A. Sauter[4]

8. CONTRACT OR GRANT NUMBER(s): 1 N00014-93-I-0131 (ONR) 2 #517 402 524 3/DAAD 4 #Sa 607/1-1 (DFG)

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742-2431

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS: Department of the Navy, Office of Naval Research, Arlington, VA 22217

12. REPORT DATE: September 1994

13. NUMBER OF PAGES: 47

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release: distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

When using the Galerkin FEM for solving the Helmholtz equation in two dimensions, the error of the corresponding solution differs substantially from the error of the best approximation, and this effect increases with higher wave number k.

In this paper we will design a Generalized Finite Element Method (GFEM) for the Helmholtz equation such that the pollution effect is minimal.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

A Generalized Finite Element Method for Solving the Helmholtz Equation in Two Dimensions with Minimal Pollution

Ivo M. Babuška Frank Ihlenburg Ellen T. Paik
Stefan A. Sauter

Abstract

When using the Galerkin FEM for solving the Helmholtz equation in two dimensions, the error of the corresponding solution differs substantially from the error of the best approximation, and this effect increases with higher wave number k.

In this paper we will design a Generalized Finite Element Method (GFEM) for the Helmholtz equation such that the pollution effect is minimal.

1. Introduction

Boundary value problems governed by the Helmholtz equation arise in many physical applications, as for example the scattering of a wave from an elastic body. For this kind of problems the computational domain consists typically of the finite domain of the elastic body coupled with the unbounded exterior domain for the scattering field. In the unbounded domain the scattered wave is described by the classical Helmholtz equation with the Laplace operator as the principal operator. In the elastic body the equation is of the same so-called Helmholtz type with the Laplace operator replaced by the elasticity operator.

In order to solve numerically such a coupled scattering problem, the finite element method is typically applied in the elastic body. For the approximation of the scattered wave, various approaches are used, such as the boundary element method, the method of infinite elements coupled with finite elements and the finite

element method where the unbounded exterior domain is replaced by a bounded artificial domain with suitable boundary conditions on the artificial boundary (see [8]).

In this paper we will address the Helmholtz problem for the Laplace operator on a bounded domain as the model problem which characterizes the behavior of the finite element method for both the Helmholtz problem of elasticity and wave scattering.

It is known and understood (see [5]) that the accuracy of the Galerkin-FEM deteriorates with increasing wave number. To be more concrete we have to introduce a norm to measure the accuracy of our FE solution. Let ϵ be the accuracy we would like to achieve. Then there exists a number of elements $n_0 = n_0(\epsilon)$ such that, in the corresponding finite element space, there is a function called "best approximation" with an error less than or equal to ϵ. Usually the FE solution needs more elements to get the same accuracy (say $n_{gal}(\epsilon)$). For standard elliptic problems, the Galerkin method is quasi-optimal, meaning that the ratio n_{gal}/n_0 is a constant. For the Helmholtz equation the situation is different. In this case the ratio n_{gal}/n_0 goes to infinity with increasing wave number. We call this non-robust behavior with respect to the wave number the "pollution effect".

A generalization of the FEM was introduced in [1] and is called Generalized FEM (GFEM). This method covers practically all modifications of the FEM which lead to a sparse system matrix. In [2] two of the authors have defined a GFEM called stabilized FEM for the Helmholtz equation in 1D with the property that $n_{stabilized}/n_0$ is a constant independent of the wave number. However, in the same paper it was proved that in the two-dimensional case there exists no GFEM such that the ratio n_{GFEM}/n_0 is bounded with respect to the wave number.

To explain the goal of this paper we consider the discretization of the Helmholtz equation separately from the discretization and incorporation of the boundary conditions, as with the finite difference method. In our paper we focus on the approximation of the Helmholtz equation in the interior of the domain by a GFEM. The approximation of the DtN mapping for the definition of the boundary condition and the effect of its discretization is not the subject of our investigation. In matrix-algebraic terms, our task is to define the interior stencil of the system matrix in such a way that under an optimal modeling of the boundary conditions the ratio $n_{stabilized}/n_0$ increases as slowly as possible.

The importance of a proper approximation of the Helmholtz operator was worked out in [2], where it was shown that, if the interior stencils lead to a

pollution, this effect cannot be countered by any discretization of the boundary conditions.

Our paper is organized as follows:

In the next section we will study a one-dimensional model problem. We will define the GFEM for the Helmholtz problem. We will show that the corresponding discretization error is directly related to the difference of the so-called discrete wave number $\tilde{k}$ with the exact one. The difference $\tilde{k}-k$ is called "phase lag". Using these results, we are able to construct a FEM with the property that $\tilde{k}-k=0$ which additionally satisfies the usual consistency conditions. We will prove that this GFEM, called "stabilized finite element method" (SFEM) has no pollution.

In Section 3 we will define the GFEM in two dimensions and a 2-D analogy to the phase difference $\tilde{k}-k$. It will turn out that in 2-D every GFEM has the phase lag. We will prove that this phase lag leads to a pollution term in the error estimates.

In Section 4 we will define and explain a measure of the approximation quality of the GFEM discretization for the Helmholtz equation.

In Section 5 we will define a GFEM called Quasi-Stabilized FEM (QSFEM) which leads to the smallest possible pollution.

In the last section we will present the results of a 2-D implementation, where we compare the quality of the quasi-stabilized FEM with the usual Galerkin FEM and a further GFEM called generalized least squares finite element method (GLS-FEM). The latter method was developed by Thompson and Pinsky (rf. [9]) based on a paper of Harari and Hughes ([4]).

2. One-dimensional model problem

In this section we will consider a one-dimensional model problem and explain why the accuracy of the Galerkin FEM deteriorates with increasing wave number k. This non-robust behavior with respect to k is called the pollution effect and will be defined formally in Definition 2.3. We will explain how the pollution effect is related to the underlying discretization method. By using that investigation we are able to construct a so-called generalized finite element method (GFEM) having no pollution.

To fix the notation let us consider the following one-dimensional model problem

$$-u''-k^2u=f \text{ in } \Omega:=(0,1) \tag{2.1}$$

with boundary conditions

$$u(0) = 0,$$

$$iku(1) + u'(1) = 0.$$

To apply the finite element method to this equation we write (2.1) in a variational formulation, seeking $u \in \mathcal{V} := \{v \in \mathcal{H}^1(\Omega) \mid v(0) = 0\}$ such that

$$a(u,v) := \int_0^1 u'\bar{v}' - k^2 u\bar{v}dx + iku(1)\bar{v}(1) = f(v) \tag{2.2}$$

is fulfilled for all $v \in \mathcal{V}$. Here and in the following we assume that f lies in the dual space $\mathcal{V}' = \mathcal{H}^{-1}(\Omega)$.

Further, let $\{x_i\}_{0\leq i\leq n}$ denote a set of grid points $0 = x_0 < x_1 < \ldots < x_n = 1$. The finite element grid τ consists of the intervals $\{[x_{m-1}, x_m]\}_{1\leq m\leq n}$. The step size h is defined by

$$h := \max_{1\leq m\leq n}(x_m - x_{m-1}).$$

We consider here only the h-version of finite elements and define $\mathcal{S}_h$ as the space of continuous functions which are linear on each interval. A study of the p-version of the Galerkin-FEM for the Helmholtz equation can be found in [7].

2.1. The GFEM for the Helmholtz equation in 1D

The Generalized Finite Element Method (GFEM) was first introduced by Babuška and Osborn (rf. [1]). The idea is to introduce local mappings which transform the usual finite element basis functions to another local basis. In the mentioned paper these local mappings were designed in such a way that the resulting method, applied to a differential equation with highly non-smooth coefficients, converges with optimal rate.

For our purpose we define the GFEM in the following algebraic way. The GFEM is a method which defines a tri-diagonal matrix $\mathbf{A} \in \mathbf{C}^{n\times n}$ and a linear mapping $\mathcal{Q} : \mathcal{H}^{-1} \to \mathbf{C}^n$. The solution of

$$\mathbf{Au} = \mathbf{b}$$

with $\mathbf{b} := \mathcal{Q}(f)$ is then identified with a finite element function by the basis representation

$$u_{fe}(x) := \sum_{m=1}^{n} \mathbf{u}_m \phi_m(x)$$

with the usual local nodal basis $\{\phi_m\}_{1\leq m\leq n}$ of $\mathcal{S}_h$. The function u_{fe} serves as an approximation of the exact solution of (2.2). How the Galerkin FEM can be written in this notation can be found in the following

Example 2.1. *The Galerkin FEM is characterized by the tri-diagonal matrix*

$$\mathbf{A}_{i,j} := a(\phi_j, \phi_i)$$

and the mapping $\mathcal{Q}$ defined by

$$\mathbf{b}_i := (\mathcal{Q}(f))_i := f(\phi_i).$$

2.2. Error analysis for the Galerkin FEM

To measure the accuracy of the FE-solution we have to introduce suitable norms. We will consider the $\mathcal{L}^2$ and $\mathcal{H}^1$-seminorm defined by

$$\|u\|_0^2 := \int_\Omega u(x)\,\bar{u}(x)\,dx$$

and

$$\|u\|_1 := \|u'\|_0.$$

From the approximation theory it is well known that for every function $u \in \mathcal{H}^2(\Omega)$ there exists $u_h \in \mathcal{S}_h$ such that

$$\|u - u_h\|_j \leq Ch^{2-j}\,\|u''\|_0$$

for $j \in \{0,1\}$. The dependency of the relative error on h and k is discussed in the following

Theorem 2.2. *Let the right-hand side f of (2.2) be in $\mathcal{L}^2(\Omega)$. We assume that the exact solution u of (2.2) is oscillating in the sense that for $0 \leq s < t \leq 2$*

$$\frac{\|u^{(t)}\|_0}{\|u^{(s)}\|_0} \leq Ck^{t-s} \tag{2.3}$$

is satisfied. For $j \in \{0,1\}$, the best approximation $u_{opt}^j \in \mathcal{S}_h$ with respect to the $\mathcal{H}^j$-seminorm is defined by

$$u_{opt}^j := \arg \inf_{u_h \in \mathcal{S}_h} \|u - u_h\|_j.$$

The function u^j_{opt} satisfies

$$e^j_{opt} := \frac{\left\|u - u^j_{opt}\right\|_j}{\|u\|_j} \leq C\,(hk)^{2-j}\,.$$

Proof. Using (2.3) we obtain

$$e^j_{opt} = \frac{\left\|u - u^j_{opt}\right\|_j}{\|u\|_j} \leq Ch^{2-j}\frac{\|u''\|_0}{\|u^{(j)}\|_0} \leq C\,(hk)^{2-j}\,.$$

■

Remark 1. *From the Theorem 2.2 we see that the accuracy of the optimal approximation depends only on the number of the elements in one wavelength of the solution, i. e., depends only on the value $k \cdot h$. To relate this value to the accuracy of the solution is a "rule of the thumb" in engineering computations.*

We say that the GFEM has the pollution effect if it is possible that e_{opt} is small but the error of the GFEM solution is arbitrarily large. The details are in the following

Definition 2.3 (pollution effect). *For $j \in \{0, 1\}$, let the error of the GFEM-solution u_{fe} be defined by*

$$e^j_{fe} := \frac{\|u - u_{fe}\|_j}{\|u\|_j}.$$

If this error can be estimated by

$$e^j_{fe} \leq C_1\,(kh)^{2-j} + C_2 k^s\,(kh)^t\,.$$

with $s > 0$ and in addition, there exists right-hand sides for problem (2.1) such that the corresponding finite element error can be estimated from below by

$$e^j_{fe} \geq C_3 k^r\,(kh)^\theta$$

with $r > 0$, then we say that the GFEM has the pollution effect.

In view of the Theorem above it is obvious that the error of the best approximation is small if kh is small, while the condition $s, r > 0$ in the definition of the pollution effect has the consequence that for sufficiently large k the quantity $k^s (kh)^t$ can be arbitrarily large, i.e. the "rule of the thumb" mentioned in the Remark 1 does not lead to an accurate solution if k is large.

The following Theorem shows that the Galerkin FEM for our model problem has the pollution effect.

Theorem 2.4. *Let the exact solution of (2.2) be oscillating in the sense of (2.3). Let us assume that our grid is uniform, which means*

$$h = x_m - x_{m-1} = x_i - x_{i-1} \quad \forall m, i \in \{1, 2, \ldots, n\}$$

Let the right-hand side of (2.2) $f \in \mathcal{L}^2(\Omega)$ and u_{gal} be the Galerkin solution. Then for $hk \leq 1$ the error estimate

$$e^1_{gal} := \frac{\|u - u_{gal}\|_1}{\|u\|_1} \leq C(kh) + C_2 k (kh)^2$$

holds. The error in the $\mathcal{L}^2$-norm can be estimated by

$$e^0_{gal} := \frac{\|u - u_{gal}\|_0}{\|u\|_0} \leq (C_3 + C_4 k)(kh)^2 .$$

Proof. In the proof of [5, Theorem 5] it was shown that

$$\|u - u_{gal}\|_1 \leq \left(C_1 h + C_2 (kh)^2\right) \|u''\|$$

holds. In conjunction with (2.3) we obtain the desired estimate of e^1_{gal}.

The $\mathcal{L}^2$-estimate was proven in [6, Theorem 4].

■

Numerical computations in [5] shows that the error estimates are optimal, i.e. there are cases where e^1_{gal} and e^0_{gal} are bounded from below by the same pollution term as from above. For a theoretical investigation see Theorem 2.6 together with Lemma 2.5.

2.3. Relation of the finite element error to the discrete wave number

In this section we will explain how the pollution effect is related to the difference of a discrete wave number and the exact one. Later, we will study this effect in two dimensions as well. We include here the one-dimensional investigation because the main ideas are more visible than in two dimensions. To avoid too many technicalities we consider the following model example with Robin boundary conditions on both sides .

$$\begin{aligned} -u'' - k^2 u &= 0 \quad \text{in } \Omega = (0,1) \\ -u'(0) - iku(0) &= -2ik, \\ u'(1) - iku(1) &= 0. \end{aligned} \tag{2.4}$$

If not stated otherwise, we assume throughout this chapter that Ω is partitioned into intervals having constant length h. It is easy to check that the exact solution of (2.4) is given by

$$u(x) = e^{ikx}.$$

We consider a GFEM of the form

$$\mathbf{Du} = \mathbf{b} \tag{2.5}$$

with the $(n+1) \times (n+1)$ matrix $\mathbf{D}$

$$\mathbf{D} = \begin{bmatrix} D_1 & N & & & & \\ N & D_2 & N & & & \\ & N & D_2 & N & & \\ & & N & \ddots & \ddots & \\ & & & \ddots & D_2 & N \\ & & & & N & D_1 \end{bmatrix}$$

and the right-hand side vector

$$\mathbf{b} = (-2ik, 0, 0, \ldots, 0)^T.$$

We assume that the elements of the matrix $\mathbf{D}$ can be expanded in a Taylor series of the form

$$D_1 = h^{-1}\left(1 - ikh + \sum_{n=1}^{p} \alpha_n (kh)^{2n} + i(kh) \sum_{n=1}^{p} \beta_n (kh)^{2n} + O\left((kh)^{2p+2}\right)\right)$$

(2.6)

$$
\begin{aligned}
D_2 &= h^{-1}\left(2+\sum_{n=1}^{p}\gamma_n\,(kh)^{2n}+O\left((kh)^{2p+2}\right)\right)\\
N &= h^{-1}\left(-1+\sum_{n=1}^{\infty}\delta_n\,(kh)^{2n}+O\left((kh)^{2p+2}\right)\right)\\
&\alpha_1+\beta_1=-\tfrac{1}{2}\qquad \gamma_1+2\delta_1=-1.
\end{aligned}
$$

These assumptions are very natural in view of the underlying equations. Some comments for the first three conditions are given later in Remark 3. The impact of the last two equations will become clear in the proof of Theorem 2.6.

The system (2.5) can be solved explicitly. For this purpose, we will use the concept of the discrete Fourier transform to solve finite difference equations. Alternatively, one could employ the theory of fundamental systems for our one-dimensional model problem. We prefer the first method, because the discrete Fourier transform can be extended straightforwardly to the higher-dimensional case. In contrast to this, the theory of fundamental systems is applicable only in the one-dimensional case because the number of homogenous solutions of a second-order differential equation in 2-D without boundary conditions is infinite.

We start by computing the discrete symbol of a difference scheme. For this purpose we introduce the discrete Fourier transform of a complex vector $\mathbf{u} = \{\mathbf{u}_m\}_{m\in\mathbf{Z}=\{\ldots,-1,0,1,\ldots\}}$ by

$$\hat{\mathbf{u}}\,(\xi) := (\mathcal{F}\mathbf{u})\,(\xi) := \sum_{m=-\infty}^{\infty}\mathbf{u}_m e^{im\xi}.$$

For a difference scheme with constant coefficients given by

$$(\mathbf{Au})_m = \sum_{l=-p}^{p}\mathbf{A}_l u_{m+l}$$

the discrete Fourier transform can be computed as

$$
\begin{aligned}
\widehat{(\mathbf{Au})}\,(\xi) &= \textstyle\sum_{m=-\infty}^{\infty}\sum_{l=-p}^{p}\mathbf{A}_l u_{m+l}e^{im\xi}\\
&= \textstyle\sum_{l=-p}^{p}\mathbf{A}_l\sum_{m=-\infty}^{\infty}u_{m+l}e^{im\xi} = \sum_{l=-p}^{p}\mathbf{A}_l e^{-il\xi}\sum_{m=-\infty}^{\infty}u_{m+l}e^{i(m+l)\xi}\\
&= \textstyle\sum_{l=-p}^{p}\mathbf{A}_l e^{-il\xi}\sum_{m=-\infty}^{\infty}u_m e^{im\xi} = \left(\sum_{l=-p}^{p}\mathbf{A}_l e^{-il\xi}\right)\hat{\mathbf{u}}\,(\xi).
\end{aligned}
$$

The function $\mathbf{a}(\xi) := \left(\sum_{l=-p}^{p} \mathbf{A}_l e^{-il\xi}\right)$ is called the discrete symbol of the difference operator $\mathbf{A}$. Let $\{\xi_l\}_{-p \le l \le p}$ denote the zeros of $\mathbf{a}$ in the interval $[-\pi, \pi[$, i.e.

$$\begin{aligned} \mathbf{a}(\xi_l) &= 0 \qquad \forall l \in \{-p, -p+1, \ldots, p\} \\ \xi_l &\in [-\pi, \pi[\,. \end{aligned}$$

By the theory of finite difference schemes it follows that the vector

$$\eta_m := \sum_{l=-p}^{p} C_l e^{i\xi_l m}, \quad m \in \mathbf{Z}$$

satisfies

$$\mathbf{A}\eta = \mathbf{0}.$$

Simple computations yield that the discrete symbol of the difference operator which corresponds to the GFEM (2.5) is given by

$$\mathbf{d}(\xi) = D_2 + 2N\cos\xi.$$

The zeros of the symbol are given by

$$\xi = \pm\tilde{k}$$

with

$$\tilde{k} = \frac{1}{h}\arccos\left(-\frac{D_2}{2N}\right). \tag{2.7}$$

The number $\tilde{k}$ in this context is called the "discrete wave number". Consequently,

$$\mathbf{u}_j = C_1 e^{i\tilde{k}j} + C_2 e^{-i\tilde{k}j}$$

satisfies equation (2.5) for indices $2 \le j \le n-1$, i.e.

$$(\mathbf{Du})_j = 0 \quad \forall j \in \{2, 3, \ldots, n-1\}.$$

The constants C_1 and C_2 are determined by the equations

$$\begin{aligned} (\mathbf{Du})_1 &= -2ik, \\ (\mathbf{Du})_n &= 0, \end{aligned}$$

in other words, by the boundary conditions. They are given explicitly by

$$C_1 = \frac{ke^{-i\tilde{k}}\left(D_1+Ne^{i\tilde{k}h}\right)}{D_1^2\sin\tilde{k}+2D_1N\sin\left(\tilde{k}(1-h)\right)+N^2\sin\left(\tilde{k}(1-2h)\right)}$$
$$C_2 = \frac{-ke^{i\tilde{k}}\left(D_1+Ne^{-i\tilde{k}h}\right)}{(D_1^2\sin\tilde{k}+2D_1N\sin\left(\tilde{k}(1-h)\right)+N^2\sin\left(\tilde{k}(1-2h)\right)} \tag{2.8}$$

with $h := 1/n$.

To summarize the explanations above, we state that the solution of our tridiagonal system of linear equations (2.5) is given by

$$\mathbf{u}_j = C_1e^{i\tilde{k}j} + C_2e^{-i\tilde{k}j}, \quad 0 \le j \le n, \tag{2.9}$$

where the discrete wave number $\tilde{k}$ is given by the zeros of the discrete symbol of the underlying difference operator and is independent of the boundary conditions. The boundary conditions then determine the constants C_1 and C_2.

Now, we will study the relative error of the GFE-solution in the $\mathcal{L}^2$-norm given by

$$e_0 := \frac{\left\|e^{ikx} - \sum_{j=0}^{n}\mathbf{u}_j\phi_j(x)\right\|}{\|e^{ikx}\|}.$$

The functions ϕ_j are the usual linear nodal basis of the finite element space $\mathcal{S}_h$.

It turns out that the error e_0 is directly related to the distance of the wave number k from the discrete wave number $\tilde{k}$. Therefore, before we start to estimate e_0, we will estimate $k - \tilde{k}$. The details are given in the following

Lemma 2.5. *Let kh be bounded and conditions (2.6) be satisfied. Then either*

$$k - \tilde{k} = 0$$

or there exist constants q_0, q_1 independent of k and h but possibly dependent on γ and δ of (2.6) such that

$$q_0k\,(kh)^{s_0} \le \left|k - \tilde{k}\right| \le q_1k\,(kh)^{s_0}$$

with $s_0 \ge 2$.

For the Galerkin-FEM the estimate above holds with $s_0 = 2$.

Proof. The discrete wave number was given by (2.7). In view of that equation we compute the quotient $\frac{D_2}{2N}$. Using (2.6) we get

$$-\frac{D_2}{2N} = -\frac{2+\sum_{n=1}^{\infty}\gamma_n (kh)^{2n}}{2\left(-1+\sum_{n=1}^{\infty}\delta_n (kh)^{2n}\right)} = 1+\left(\frac{1}{2}\gamma_1+\delta_1\right)(kh)^2+\sum_{s=2}^{\infty}\rho_s (kh)^{2s}.$$

The modified wave number, given by

$$\tilde{k} = \frac{1}{h}\arccos\left(1+\left(\frac{1}{2}\gamma_1+\delta_1\right)(kh)^2+\sum_{s=2}^{\infty}\rho_s (kh)^{2s}\right),$$

can be expanded about $kh = 0$ as

$$\tilde{k} = \frac{1}{h}\left(\sqrt{(-\gamma_1-2\delta_1)}kh+\sum_{s=1}^{\infty}\iota_s (kh)^{2s+1}\right).$$

Now it is clear why we imposed the fourth condition in (2.6). Under this assumption, the term $\sqrt{-(\gamma_1+2\delta_1)} = 1$ and the discrete wave number converges towards k as $h \to 0$. The equation above can then be written in the form

$$\tilde{k} - k = O\left(k\,(kh)^{s_0}\right)$$

with even $s_0 \geq 2$. Only in the case of $\iota_s = 0$ for all $s \geq 1$ we obtain

$$\tilde{k} - k = 0.$$

In [5] it was shown that for the Galerkin method $\tilde{k} = k + O\left(k\,(kh)^2\right)$ holds if kh is bounded. ■

In the sequel the number s_0 is given by the Lemma above.

Now, we will estimate the relative $\mathcal{L}^2$-error e_0 from above and below. The details are in the following

Theorem 2.6. *Let us assume that the considered GFEM has the property $k \neq \tilde{k}$. This means that s_0, defined above is finite. Let kh and $k\,(kh)^{s_0}$ be bounded and k sufficiently large.*

Then, the error e_0 of the GFE-approximation of the solution of (2.4), namely e^{ikx}, can be estimated by

$$\tilde{c}\left|k-\tilde{k}\right| \leq e_0 \leq C\,(kh)^2 + \tilde{C}\left|k-\tilde{k}\right|.$$

Proof. For the following analysis, we will use the interpolant u_{int}^{ω} of the function $e^{i\omega x}$, i.e.

$$u_{int}^{\omega}(x) = \sum_{j=0}^{n} e^{i\omega jh}\phi_j(x).$$

Using the fact that $\left\|e^{ikx}\right\| = 1$, the error e_0 can be estimated from above by

$$\begin{aligned} e_0 &= \left\|e^{ikx} - e^{i\tilde{k}x} + e^{i\tilde{k}x} - u_{int}^{\tilde{k}} + u_{int}^{\tilde{k}} - \textstyle\sum_{j=0}^{n} \mathbf{u}_j\phi_j(x)\right\| \\ &\leq \left\|e^{ikx} - e^{i\tilde{k}x}\right\| + \left\|e^{i\tilde{k}x} - u_{int}^{\tilde{k}}\right\| + \left\|u_{int}^{\tilde{k}} - \textstyle\sum_{j=0}^{n} \mathbf{u}_j\phi_j(x)\right\|. \end{aligned} \tag{2.10}$$

We will now estimate the three terms on the right-hand side above separately. Considering the first term we get:

$$\begin{aligned} \left\|e^{ikx} - e^{i\tilde{k}x}\right\|^2 &= \textstyle\int_0^1 \left(e^{ikx} - e^{i\tilde{k}x}\right)\left(e^{-ikx} - e^{-i\tilde{k}x}\right)dx \\ &= 2\left(1 - \tfrac{\sin\left(k-\tilde{k}\right)}{k-\tilde{k}}\right). \end{aligned} \tag{2.11}$$

By our assumption we know that $k - \tilde{k}$ is bounded, thus the Taylor expansion about $k - \tilde{k} = 0$ results in

$$\left\|e^{ikx} - e^{i\tilde{k}x}\right\| \leq \frac{\left|k-\tilde{k}\right|}{\sqrt{3}} + O\left(\left|k-\tilde{k}\right|^3\right).$$

The second term on the right-hand side of (2.10) can be estimated by using a standard interpolation argument

$$\left\|e^{i\tilde{k}x} - u_{int}^{\tilde{k}}\right\| \leq Ch^2\left\|\left(e^{i\tilde{k}x}\right)''\right\| = C\left(h\tilde{k}\right)^2.$$

Using $\tilde{k} = k + O\left(k\,(kh)^{s_0}\right)$ (see Lemma 2.5), we obtain

$$\left\|e^{i\tilde{k}x} - u_{int}^{\tilde{k}}\right\| \leq C\,(hk)^2.$$

For the last term in (2.10) we proceed as follows:

$$\left\|u_{int}^{\tilde{k}} - \sum_{j=0}^{n} \mathbf{u}_j\phi_j(x)\right\|^2 = \sum_{l,m=0}^{n} \left(\mathbf{u}_{int}^{\tilde{k}} - \mathbf{u}\right)_l \mathbf{M}_{l,m}\overline{\left(\mathbf{u}_{int}^{\tilde{k}} - \mathbf{u}\right)}_m$$

with the mass matrix $\mathbf{M}$ defined by

$$\mathbf{M}_{i,j} := \int_0^1 \phi_i(x)\,\phi_j(x)\,dx$$

and $\left(\mathbf{u}_{int}^{\tilde{k}}\right)_m := u_{int}^{\tilde{k}}(hm)$. It is well-known that if the FE-space corresponds to a uniform grid, the $\mathcal{L}^2$-norm is equivalent to the weighted Euclidean norm, resulting in

$$\left\| u_{int}^{\tilde{k}} - \sum_{j=0}^{n} \mathbf{u}_j \phi_j(x) \right\|^2 \leq h \sum_{m=0}^{n} \left| \left(\mathbf{u}_{int}^{\tilde{k}} - \mathbf{u}\right)_m \right|^2 =: \left\| \mathbf{u}_{int}^{\tilde{k}} - \mathbf{u} \right\|_{l^2}^2 .$$

The norm $\|\cdot\|_{l^2}$ is called the l^2-norm. For a vector $\mathbf{u}$, we introduce the convention

$$\|\mathbf{u}_m\|_{l^2}^2 := h \sum_{m=0}^{n} |\mathbf{u}_m|^2 .$$

Using the definition of $\mathbf{u}_{int}^{\tilde{k}}$ and (2.9) we obtain

$$\left\| u_{int}^{\tilde{k}} - \sum_{j=0}^{n} \mathbf{u}_j \phi_j(x) \right\| \leq \left\| e^{ikjh} - C_1 e^{i\tilde{k}jh} - C_2 e^{-i\tilde{k}jh} \right\|_{l^2} \leq |1 - C_1| \left\| e^{i\tilde{k}jh} \right\|_{l^2} + |C_2| \left\| e^{-i\tilde{k}jh} \right\|_{l^2} .$$

Direct calculations yield

$$\left\| e^{i\tilde{k}j} \right\|_{l^2}^2 = h \sum_{j=0}^{n} \left| e^{i\tilde{k}jh} \right|^2 = h \sum_{j=0}^{n} 1 = (n+1)\,h \leq 2,$$

thus,

$$\left\| u_{int}^{\tilde{k}} - \sum_{j=0}^{n} \mathbf{u}_j \phi_j(x) \right\| \leq 2\left(|1 - C_1| + |C_2|\right).$$

Let us first estimate the constant C_2 of (2.8):

$$C_2 = \frac{-ke^{i\tilde{k}}\left(D_1 + Ne^{-i\tilde{k}h}\right)}{D_1^2 \sin\tilde{k} + 2D_1 N \sin\left(\tilde{k}\,(1-h)\right) + N^2 \sin\left(\tilde{k}\,(1-2h)\right)}.$$

Replacing $\tilde{k}$ by $k + \epsilon$ and inserting the formulae for D_1 and N (cf. 2.6) into the definition of C_2 and expanding C_2 as a Taylor series about $\epsilon = 0$ and $hk = 0$,we get

$$C_2 = \left| \frac{1 + 2\,(\alpha_1 + \beta_1)}{4} \right| kh + C\,(kh)^2 .$$

In view of this representation it is clear why we imposed the fourth condition in (2.6). The constant C_2 then can be estimated by

$$C_2 \leq C\,(kh)^2\,. \tag{2.12}$$

Applied to $|1 - C_1|$, the same arguments results in

$$|1 - C_1| \leq C\,(kh)^2\,. \tag{2.13}$$

Combining all estimates above we conclude that the inequality

$$\left\| e^{ikx} - \sum_{j=0}^{n} \left(C_1 e^{i\tilde{k}jh} + C_2 e^{-i\tilde{k}jh} \right) \phi_j\,(x) \right\| \leq C\,(kh)^2 + \tilde{C}\left| k - \tilde{k} \right|$$

is fulfilled, completing the proof of the upper estimate.

For the lower estimate we proceed in the following way. The error e_0 can be written in the form

$$e_0 = \left\| \left(e^{ikx} - C_1 e^{i\tilde{k}x} - C_2 e^{-i\tilde{k}x} \right) + \left(C_1 e^{i\tilde{k}x} + C_2 e^{-i\tilde{k}x} - \sum_{j=0}^{n} \mathbf{u}_j \phi_j\,(x) \right) \right\|.$$

By definition we have that $\sum_{j=0}^{n} \mathbf{u}_j \phi_j\,(x)$ is the interpolant of $C_1 e^{i\tilde{k}x} + C_2 e^{-i\tilde{k}x}$. Therefore, we know that

$$\left\| C_1 e^{i\tilde{k}x} + C_2 e^{-i\tilde{k}x} - \sum_{j=0}^{n} \mathbf{u}_j \phi_j\,(x) \right\| \leq Ch^2 \left\| \left(C_1 e^{i\tilde{k}x} + C_2 e^{-i\tilde{k}x} \right)'' \right\| \leq C\,(kh)^2\,(|C_1| + |C_2|)\,.$$

Using (2.12,2.13) we obtain that

$$\left\| C_1 e^{i\tilde{k}x} + C_2 e^{-i\tilde{k}x} - \sum_{j=0}^{n} \mathbf{u}_j \phi_j\,(x) \right\| \leq C_{int}\,(kh)^2\,.$$

Thus, e_0 can be estimated by

$$e_0 \geq \left\| e^{ikx} - C_1 e^{i\tilde{k}x} - C_2 e^{-i\tilde{k}x} \right\| - C_{int}\,(kh)^2 \geq \min_{u,v \in \mathbf{C}} \left\| e^{ikx} - u e^{i\tilde{k}x} - v e^{-i\tilde{k}x} \right\| - C_{int}\,(kh)^2\,.$$

Let $L := \left\lfloor \frac{\tilde{k}}{2\pi} \right\rfloor \cdot \frac{2\pi}{k}$ where $\lfloor q \rfloor$ denotes the largest integer less than or equal to q. Let $\|\cdot\|_L$ be defined by

$$\|u\|_L^2 := \int_0^L u\,(x)\,\bar{u}\,(x)\,dx.$$

Using this norm, e_0 can be estimated by

$$e_0 \geq \min_{u,v \in \mathbf{C}} \left\| e^{ikx} - ue^{i\tilde{k}x} - ve^{-i\tilde{k}x} \right\|_L - C_{int} (kh)^2 .$$

The computation of this minimum is easy in view of the orthogonality of $e^{i\tilde{k}x}$ and $e^{-i\tilde{k}x}$ on $(0, L)$. Introducing $\epsilon = k - \tilde{k}$, we get that the minimum is achieved for

$$\begin{aligned} u_0 &= \frac{1}{L} \int_0^L e^{i(k-\tilde{k})x} dx = \frac{e^{i\epsilon L} - 1}{i\epsilon L} \\ v_0 &= \frac{1}{L} \int_0^L e^{i(k+\tilde{k})x} dx = \frac{e^{i(2\tilde{k}+\epsilon)L} - 1}{i\left(2\tilde{k}+\epsilon\right)L} = \frac{e^{i\epsilon L} - 1}{i\left(2\tilde{k}+\epsilon\right)L}. \end{aligned}$$

We assumed that k is sufficiently large, therefore $L = \left\lfloor \frac{\tilde{k}}{2\pi} \right\rfloor \cdot \frac{2\pi}{\tilde{k}} \geq C \left\lfloor \frac{k}{2\pi} \right\rfloor \cdot \frac{2\pi}{k}$ is bounded away from zero and the minimum above can be computed to

$$\begin{aligned} \min_{u,v \in \mathbf{C}} \left\| e^{ikx} - ue^{i\tilde{k}x} - ve^{-i\tilde{k}x} \right\|_L &= \sqrt{1 - |u_0|^2 - |v_0|^2} \\ &= \sqrt{1 - 2\left(1 - \cos\left(\epsilon L\right)\right)\left(\frac{1}{(\epsilon L)^2} + \frac{1}{\left(\left(2\tilde{k}+\epsilon\right)L\right)^2}\right)} \\ &\geq C\epsilon. \end{aligned}$$

Combining the above estimates we obtain that

$$e_0 \geq C\epsilon - C_{int} (kh)^2 .$$

For sufficiently large k the term $C_{int} (kh)^2$ becomes negligible compared to $\epsilon = O\left(k (kh)^{s_0}\right)$, resulting in

$$e_0 \geq C\epsilon = C \left| k - \tilde{k} \right|$$

completing the proof.

■

2.4. A stabilized finite element method

In view of Theorem 2.6 it is clear that a GFEM with no pollution term in the error estimates must satisfy

$$k - \tilde{k} = 0.$$

This is equivalent to

$$\frac{D_2}{N} = -2\cos(kh). \tag{2.14}$$

Of course, the analysis of the model problem (2.4) guarantees that the arising so-called "stabilized finite element method" (SFEM) has no pollution only for this special example. However, two of the authors considered the more general problem (2.1) and showed that if (2.14) is fulfilled, it is always possible to discretize the boundary conditions and the inhomogeneous right-hand side in such a way that the GFEM has no pollution. The details are in the following

Theorem 2.7 (stabilized finite element method). *Let us consider the problem (2.1). Let the interval* $(0,1)$ *be partitioned using the grid points* $0 = x_0 < x_1 < \ldots < x_n = 1$. *Here, we do not require a uniform grid. Let the* $n \times n$ *system matrix* $\mathbf{D}^{stab}$ *be given by*

$$\mathbf{D}^{stab}_{i,j} = \frac{k^2 h}{2\tan\frac{kh}{2}} \begin{cases} \frac{\sin(k(x_{i+1}-x_{i-1}))}{\sin(k(x_{i+1}-x_i))\sin(k(x_i-x_{i-1}))} & \text{if } i=j<n, \\ \frac{e^{-ik(x_n-x_{n-1})}}{\sin(k(x_n-x_{n-1}))} & \text{if } i=j=n, \\ -\frac{1}{\sin(k\lvert x_i-x_j\rvert)} & \text{if } \lvert j-i\rvert=1, \\ 0 & \text{otherwise} \end{cases}$$

and the mapping $\mathcal{Q}^{stab}$ *by*

$$\left(\mathcal{Q}^{stab}\right)_i = \frac{h}{2\tan\frac{kh}{2}} \sum_{m=i}^{\min(i+1,n)} \frac{\tan\left(k\frac{x_m-x_{m-1}}{2}\right)}{x_m - x_{m-1}} \frac{\int_{x_{m-1}}^{x_m} f(x)\,dx}{(x_m - x_{m-1})}.$$

The corresponding GFE-solution $\mathbf{u}$ *is defined by*

$$\mathbf{Du} = \mathbf{f},$$

while u denotes the exact solution of (2.1).

Under the assumption that the right-hand side of (2.1) $f \in \mathcal{H}^1(0,1)$, the error estimate

$$\frac{\left\| u - \sum_{j=1}^{n} \mathbf{u}_j \phi_j \right\|_1}{\|u\|_1} \leq C(kh)$$

is satisfied. Obviously, this so-called stabilized finite element method (SFEM) has no pollution.

Proof. The proof of this Theorem can be found in [2, Theorem 2.3].

Remark 2. *In the case of a uniform mesh, for the SFEM, the quotient D_2/N satisfies*

$$\frac{D_2}{N} = -2\cos(kh).$$

3. The GFEM to the Helmholtz equation in two dimensions

In this section we consider the Helmholtz equation on a square of side length $2L$

$$-\Delta u - k^2 u = f \quad \text{in } \Omega = (-L, L)^2 \tag{3.1}$$

with boundary conditions

$$\alpha_s iku + \beta_s \frac{\partial u}{\partial n} = g_s \text{ on } \Gamma_s \text{ for } s \in \{1,2,3,4\}. \tag{3.2}$$

The boundary pieces Γ_s are depicted in figure (3.1) where the symbol $\frac{\partial}{\partial n}$ means "outside normal derivative".

Let Ω be partitioned into squares of side length $h = \frac{2L}{n-1}$ forming the finite element grid τ. The space of finite element functions $\mathcal{S}_h$ is defined by continuous, piecewise-bilinear functions corresponding to the grid τ. The grid points are defined by $x_{s,t} := h(t-1, s-1)^T$. The local basis functions are denoted by $\{\phi_{s,t}\}_{1 \leq s,t \leq n}$, and satisfy the relation

$$\phi_{s,t}(x_{s',t'}) = \begin{cases} 1 & \text{if } (s,t) = (s',t') \\ 0 & \text{otherwise.} \end{cases}$$

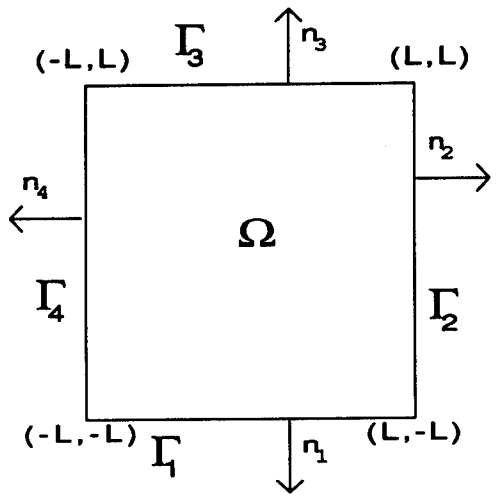


Figure 3.1: Domain Ω with boundary and normal directions

A finite element function $u \in \mathcal{S}_h$ can be identified with a vector $\{\mathbf{u}_{s,t}\}_{1\leq s,t\leq n}$ by the basis representation

$$u(x) = \sum_{s,t=1}^{n} \mathbf{u}_{s,t}\phi_{s,t}(x). \tag{3.3}$$

The GFEM is a method which defines an $n^2 \times n^2$ matrix $\mathbf{A}$ and a linear mapping $\mathcal{Q}$ which maps the right-hand sides of (3.1,3.2) onto the vector of the right-hand side $\mathbf{b}$, i.e.

$$\mathbf{b} := \mathcal{Q}(g, f).$$

The solution of the linear system

$$\mathbf{Au} = \mathbf{b}$$

is identified with a finite element function by (3.3) which serves as an approximation of the exact solution of (3.1,3.2).

The dimension of the system matrix is n^2 thus, each grid point $x_{s,t}$ can be associated with one equation. If the grid point $x_{s,t}$ is an interior grid point, i.e. not lying on the boundary, then the corresponding equation can be written in the

form

$$\begin{array}{lllll} & A_{s,t}^{-1,1}\mathbf{u}_{s-1,t+1} & +A_{s,t}^{0,1}\mathbf{u}_{s,t+1} & +A_{s,t}^{1,1}\mathbf{u}_{s+1,t+1} & \\ + & A_{s,t}^{-1,0}\mathbf{u}_{s-1,t} & +A_{s,t}^{0,0}\mathbf{u}_{s,t} & +A_{s,t}^{1,0}\mathbf{u}_{s+1,t} & \\ + & A_{s,t}^{-1,-1}\mathbf{u}_{s-1,t-1} & +A_{s,t}^{0,-1}\mathbf{u}_{s,t-1} & +A_{s,t}^{1,-1}\mathbf{u}_{s+1,t-1} & = \quad b_{s,t}. \end{array}$$

If the grid point is lying on the boundary Γ, then those elements of $A_{s,t}^{(\cdot,\cdot)}$ which are multiplied with values of $u_{p,q}$ with $h\,(q-1, p-1)^T \notin \bar{\Omega}$ have to be set to zero. This means that the system matrix of a GFEM is set up by defining all interior "stencils"

$$A_{s,t}^{interior} := \begin{bmatrix} A_{s,t}^{-1,1} & A_{s,t}^{0,1} & A_{s,t}^{1,1} \\ A_{s,t}^{-1,0} & A_{s,t}^{0,0} & A_{s,t}^{1,0} \\ A_{s,t}^{-1,-1} & A_{s,t}^{0,-1} & A_{s,t}^{1,-1} \end{bmatrix}$$

all edge stencils

$$A_{s,t}^{edge} = \begin{bmatrix} A_{s,t}^{-1,1} & A_{s,t}^{0,1} & A_{s,t}^{1,1} \\ A_{s,t}^{-1,0} & A_{s,t}^{0,0} & A_{s,t}^{1,0} \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} A_{s,t}^{-1,1} & A_{s,t}^{0,1} & 0 \\ A_{s,t}^{-1,0} & A_{s,t}^{0,0} & 0 \\ A_{s,t}^{-1,-1} & A_{s,t}^{0,-1} & 0 \end{bmatrix}, \quad etc.$$

and all corner stencils

$$A_{s,t}^{corner} = \begin{bmatrix} 0 & A_{s,t}^{0,1} & A_{s,t}^{1,1} \\ 0 & A_{s,t}^{0,0} & A_{s,t}^{1,0} \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} A_{s,t}^{-1,1} & A_{s,t}^{0,1} & 0 \\ A_{s,t}^{-1,0} & A_{s,t}^{0,0} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad etc.$$

The quality of our GFEM approximation depends on the coefficients $A_{s,t}^{(\cdot,\cdot)}$ and the definition of the right-hand side vector. We impose the following restrictions on the stencils $A_{s,t}$ and mapping $\mathcal{Q}$.

A1 The interior stencils $A_{s,t}^{interior}$, which depend on k and h, satisfy the following symmetry condition

$$A_{s,t}^{interior} := \begin{bmatrix} A_2 & A_1 & A_2 \\ A_1 & A_0 & A_1 \\ A_2 & A_1 & A_2 \end{bmatrix} \tag{3.4}$$

and have the same values for all s and t.

A2 Let us assume that the right-hand side f of (3.1) is zero. Then, the linear operator $\mathcal{Q}$ is local in such a way that if $x_{s,t}$ is a grid point satisfying $\operatorname{dist}(x_{s,t}, \Gamma) \geq 2h$ then the corresponding entry $\mathbf{b}_{s,t}$ of the right-hand side vector is zero.

A3 We assume that the interior stencils of the finite element matrices $A^{interior}$ can be expanded as a Taylor series of the form

(i) $A_0 = \sum_{m=0}^{\infty} (A_0)_m \alpha^{2m}$,
(ii) $A_1 = \sum_{m=0}^{\infty} (A_1)_m \alpha^{2m}$,
(iii) $A_2 = \sum_{m=0}^{\infty} (A_2)_m \alpha^{2m}$,
with $\alpha = kh$ and $(A_t)_m$ independent of k and h for all $t \in \{0,1,2\}, m \in \{0,1,2,\ldots\}$.

A4 We assume that the principal part of A, i.e.

$$A_{principal} := \begin{bmatrix} (A_2)_0 & (A_1)_0 & (A_2)_0 \\ (A_1)_0 & (A_0)_0 & (A_1)_0 \\ (A_2)_0 & (A_1)_0 & (A_2)_0 \end{bmatrix}$$

is an approximation of the principal part $a_0(u,v) = \int_\Omega \langle \nabla u, \nabla \bar{v} \rangle \, dx$ of consistency order 2, implying

$$\begin{aligned} (A_0)_0 &> 0, \\ (A_0)_0 + 4((A_1)_0 + (A_2)_0) &= 0, \\ -(A_1)_0 - 2(A_2)_0 &= 1. \end{aligned} \tag{3.5}$$

These restrictions are very natural considering linear finite elements. Some comments are given in the following

Remark 3. *Condition A1 reflects to the rotational and translational symmetry of the Helmholtz equation and the mesh τ.*

Condition A2 reflects the fact that the discretization of the boundary conditions has local influence to the right-hand side vector.

Condition A3 reflects the fact that the Laplacian and the identity operator are of even order.

Condition A4 is the usual consistency condition if "$-\Delta$" is discretized by linear elements.

4. Approximation of the Helmholtz equation by the GFEM

In this section we will investigate the dependency of the accuracy of the GFEM on the matrix stencils. To measure the accuracy, or the difference between the GFE solution and the exact solution, we introduce a weighted $\mathcal{L}^2$-norm defined by

$$\|v\|_-^2 := \int_\Omega \frac{v(x)\,\bar{v}(x)}{1+\|x\|^2} dx.$$

This weighted norm reflects the fact that in this paper our aim is not the modeling of the DtN boundary conditions and their discretization but to discretize the Helmholtz operator in the interior of the domain in an optimal way.

To specify the quality of our GFE discretization we will use some tools of the theory of the (integral) Fourier transform and the discrete Fourier transform. We give here only a short summary of the theory presented in [2]. The symbol of the Helmholtz operator is given by

$$\sigma(\xi) = \|\xi\|^2 - k^2$$

where $\xi \in \mathbf{R}^2$ and $\|\xi\|^2 := \xi_1^2 + \xi_2^2$.

In Condition A1 of the previous section we assumed that the interior stencils of the GFE-matrix have constant coefficients, i.e.

$$A_{s,t}^{interior} = \begin{bmatrix} A_2 & A_1 & A_2 \\ A_1 & A_0 & A_1 \\ A_2 & A_1 & A_2 \end{bmatrix}$$

where A_t only depends on k and h. The discrete symbol of the corresponding difference operator is given by

$$\sigma_{stencil}(\xi) := A_0 + 2A_1(\cos\xi_1 + \cos\xi_2) + 4A_2\cos\xi_1\cos\xi_2.$$

Let $\mathcal{N}_{kh}$ be defined as the scaled roots of the operator symbol σ:

$$\mathcal{N}_{kh} := \left\{\xi \in \mathbf{R}^2 \mid \sigma\left(h^{-1}\xi\right) = 0\right\};$$

in other words $\mathcal{N}_{kh}$ is a circle centered at the origin with radius kh.

Further, let $\mathcal{N}_{stencil}$ be defined as those roots of $\sigma_{stencil}$ lying in the square $(-kh-\epsilon, kh+\epsilon) \times (-kh-\epsilon, kh+\epsilon)$ where $\epsilon > 0$ has to be chosen sufficiently

small such that $\mathcal{N}_{kh}$ is a simple connected line. The maximal distance between these curves defined by

$$\mathcal{D}(stencil) := \mathcal{D}(N_{kh}, \mathcal{N}_{stencil}) := \max_{t\in[-\pi,\pi[} \min_{\xi\in\mathcal{N}_{stencil}} \left\| kh \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} - \xi \right\| \tag{4.1}$$

can be considered as a measure of the approximation quality of the GFEM for the Helmholtz equation in the interior by the GFEM difference operator. The details can be found in the following

Theorem 4.1. *Let $A^{interior}$ be the interior stencil of a GFEM to solve the Helmholtz problem*

$$-\Delta u - k^2 u = f \text{ in } \Omega_L := (-L, L) \times (-L, L) \tag{4.2}$$

with boundary conditions

$$Bu = g \tag{4.3}$$

which should imply existence and uniqueness of the solution.

a) Then there exists $L < \infty$, a right-hand side f, and boundary data g such that the error of the GFE-solution u_{fe} can be estimated from below by

$$\|u - u_{fe}\|_{-} \geq C_1 \sqrt{\frac{\mathcal{D}(stencil)}{h}}.$$

On the other hand there exists a function u_{opt} in the finite element space which satisfies

$$\|u - u_{opt}\|_{-} \leq C_2 k^2 h^2,$$

where the constants C_1 and C_2 are independent of k and h.

b) The function $\mathcal{D}(stencil)$ can be expanded accordingly:

$$\mathcal{D}(stencil) = r_{l_0} (kh)^{2l_0+1} + O(kh)^{2l_0+3}$$

with constants r_l depending on the stencil coefficients $(A_t)_m$ for $t \in \{0, 1, 2\}, m \in \mathbf{N}_0$ (cf. Condition A3) but not on k and h. Further, $1 \leq l_0 < \infty$ and the coefficient $r_{l_0} \neq 0$.

c) Asymptotically (i.e. for sufficiently small kh) the error $\|u - u_{fe}\|_{-}$ can be estimated by

$$\|u - u_{fe}\|_{-} \geq C_1 C_{stencil} \sqrt{k} (kh)^{l_0}.$$

Consequently, if kh is small enough, the error of the best approximation could be small, but for sufficiently large k the error of the GFE-solution could be arbitrarily large.

Proof. The proof of this Theorem is given in [2, chap. 3]. ■

The theorem can be interpreted as follows. Let a GFEM be given. Then the following situation can arise for given right-hand side f and boundary data g. We want to compute the solution of (4.2,4.3) with an accuracy of ϵ. Then by the approximation property of our finite element space $\mathcal{S}_h$ we know that if the number of elements is larger than n_0 there exists a function $u_{opt} \in \mathcal{S}_h$ satisfying

$$\|u - u_{opt}\| \leq \epsilon. \tag{4.4}$$

Then, the number of elements n_{GFEM} to guarantee that the GFE-solution also fulfills (4.4) has the property that the ratio n_{GFEM}/n_0 behaves asymptotically like

$$\frac{n_{GFEM}}{n_0} \geq C \sqrt[l_0]{k}\epsilon^{1-\frac{2}{l_0}}.$$

Obviously the ratio $\frac{n_{GFEM}}{n_0}$ goes to infinity with increasing k. We conclude that the function $\mathcal{D}(stencil)$ is a well-suited measure of the approximation quality of the GFE discretization to solutions for the Helmholtz equation.

5. A generalized finite element method having minimal pollution

Based on the theoretical results of the previous section we are now able to construct a GFEM having minimal pollution. To be more concrete we will define an interior stencil such that $\mathcal{D}(stencil)$ is asymptotically minimal.

By Assumption A3 of Section 2, $A_0 \neq 0$, therefore the quotients a_1 and a_2 are well-defined by

$$a_1 = 4\frac{A_1}{A_0}, \qquad a_2 = 4\frac{A_2}{A_0}.$$

Using A4, a_1 and a_2 can be expanded as

$$a_1 = \sum_{m=0}^{\infty} \lambda_m (kh)^{2m}$$

and

$$a_2 = \sum_{m=0}^{\infty} \iota_m (kh)^{2m}. \tag{5.1}$$

Note that condition (3.5) implies that

$$\begin{aligned} 1+\lambda_0+\iota_0 &= 0 \\ \lambda_0+2\iota_0 &\neq 0. \end{aligned} \tag{5.2}$$

In view of the definition of $\mathcal{D}$ (*stencil*) (cf. 4.1) we introduce the function *dist* $(-\pi,\pi) \to \mathbf{R}_0^+$ by

$$dist\,(\beta) := \min_{\xi \in \mathcal{N}_{stencil}} \left\| kh \begin{pmatrix} \cos\beta \\ \sin\beta \end{pmatrix} - \xi \right\|.$$

The function *dist* can be expressed explicitly by the expansion

$$dist\,(\beta) := \left| \sum_{m=1}^{\infty} r_m\,(\beta)\,(kh)^{2m+1} \right|.$$

Before we define the coefficients $r_m = r_m\,(\beta)$ we introduce κ_n, τ_n and $\rho_{2n,m}$ by

$$\kappa_n = \frac{\left(-\cos^2\beta\right)^n}{(2n)!} \qquad \tau_n = \frac{\left(-\sin^2\beta\right)^n}{(2n)!} \qquad (\kappa \star \tau)_n = \textstyle\sum_{m=0}^{n} \kappa_m \tau_{n-m}$$

$$\rho_{2n,m} = \begin{cases} \delta_{0,m} & \text{if } n = 0, \\ \left(\underbrace{\tau \star \tau \star \ldots \star \tau}_{2n-fold\ convolution} \right)_m & \text{otherwise} \end{cases}$$

with $\delta_{n,m}$ denoting the Kronecker delta.

Formally we set $r_0 \equiv 1$. Then, all other coefficients $r_m\,(\beta)$ are given by the condition

$$\sum_{n=0}^{l} \sum_{m=0}^{l-n} \rho_{2n,m} \left(\lambda_{l-n-m} \frac{\kappa_n + \tau_n}{2} + \iota_{l-n-m} \left(\kappa \star \tau\right)_n \right) = 0 \quad \forall l \geq 1. \tag{5.3}$$

We state that condition (5.3) can be written in the form

$$-\frac{r_{l-1}\,(\beta)}{2}\,(\lambda_0 + 2\iota_0) + l.o.t. = 0. \tag{5.4}$$

The abbreviation *l.o.t.* denotes the remaining sum of (5.3) containing only functions r_j with $j < l-1$. In view of (5.2), relation (5.4) serves as a recursive formula

for the functions r_j. The proofs and development of these formulae can be found in [2, Appendix].

In the mentioned paper it was further proved that, for bounded $kh < \alpha_0$,

$$\mathcal{D}(stencil) := \max_{-\pi \le \beta < \pi} dist(\beta) \le C (kh)^{2l_0+1} \left| \max_{-\pi \le \beta < \pi} r_{l_0}(\beta) \right|, \tag{5.5}$$

where the largest possible value of l_0 is 3.

To summarize this section, we state that the quality measure $\mathcal{D}(stencil)$ can be explicitly computed by formulae (5.3) and (5.5) for each generalized finite element method.

By the consideration above it is clear that an asymptotically optimal interior stencil has to be designed such that $r_1(\beta) \equiv r_2(\beta) \equiv 0$. According to condition A3 (cf. Section 3), we had assumed that the interior stencil $A_{interior}$ can be written in the form

$$A^{interior} = \begin{bmatrix} A_2 & A_1 & A_2 \\ A_1 & A_0 & A_1 \\ A_2 & A_1 & A_2 \end{bmatrix} = \sum_{m=0}^{\infty} (kh)^2 A_m^{interior} = \sum_{m=0}^{\infty} (kh)^{2m} \begin{bmatrix} A_{m,2} & A_{m,1} & A_{m,2} \\ A_{m,1} & A_{m,0} & A_{m,1} \\ A_{m,2} & A_{m,1} & A_{m,2} \end{bmatrix}, \tag{5.6}$$

where the interior stencils $A_m^{interior}$ in the expansion above are independent of k and h,i.e., $A_{m,t}$ are in general complex numbers. In the following we will use $\alpha := kh$.

We define the interior stencil $A_{QSFEM}^{interior}$ of the *quasi−stabilized FEM (QSFEM)* by

$$\begin{aligned} A_0 &= 4 \\ A_1 &= 2\frac{c_1(\alpha)s_1(\alpha) - c_2(\alpha)s_2(\alpha)}{c_2(\alpha)s_2(\alpha)(c_1(\alpha)+s_1(\alpha)) - c_1(\alpha)s_1(\alpha)(c_2(\alpha)+s_2(\alpha))} \\ A_2 &= \frac{c_2(\alpha)+s_2(\alpha)-c_1(\alpha)-s_1(\alpha)}{c_2(\alpha)s_2(\alpha)(c_1(\alpha)+s_1(\alpha)) - c_1(\alpha)s_1(\alpha)(c_2(\alpha)+s_2(\alpha))}, \end{aligned} \tag{5.7}$$

while the auxiliary functions c_1, s_1, c_2, and s_2 are defined by

$$c_1(\alpha) := \cos\left(\alpha \cos\tfrac{\pi}{16}\right) \qquad s_1(\alpha) := \cos\left(\alpha \sin\tfrac{\pi}{16}\right)$$

$$c_2(\alpha) := \cos\left(\alpha \cos\tfrac{3\pi}{16}\right) \qquad s_2(\alpha) := \cos\left(\alpha \sin\tfrac{3\pi}{16}\right).$$

The function A_1 and A_2 be expanded according to (5.6) because cos is an even function. The first terms of this expansion, i.e. the stencils $A_m^{interior}$ for $m \le 4$,

are given by

$$\mathcal{A}_0^{interior} = \begin{bmatrix} -\frac{1}{5} & -\frac{4}{5} & -\frac{1}{5} \\ -\frac{4}{5} & 4 & -\frac{4}{5} \\ -\frac{1}{5} & -\frac{4}{5} & -\frac{1}{5} \end{bmatrix} \quad \mathcal{A}_1^{interior} = \begin{bmatrix} -\frac{17}{250} & -\frac{29}{125} & -\frac{17}{250} \\ -\frac{29}{125} & 0 & -\frac{29}{125} \\ -\frac{17}{250} & -\frac{29}{125} & -\frac{17}{250} \end{bmatrix}$$

$$\mathcal{A}_2^{interior} = \begin{bmatrix} -\frac{801}{50000} & -\frac{2549}{50000} & -\frac{801}{50000} \\ -\frac{2549}{50000} & 0 & -\frac{2549}{50000} \\ -\frac{801}{50000} & -\frac{2549}{50000} & -\frac{801}{50000} \end{bmatrix}$$

$$\mathcal{A}_3^{interior} = \begin{bmatrix} -\frac{76313}{22500000} & -\frac{473849}{45000000} & -\frac{76313}{22500000} \\ -\frac{473849}{45000000} & 0 & -\frac{473849}{45000000} \\ -\frac{76313}{22500000} & -\frac{473849}{45000000} & -\frac{76313}{22500000} \end{bmatrix}$$

$$A_4^{interior} = \begin{bmatrix} -\frac{826713271}{1188000000000} & -\frac{5094901033}{2376000000000} & -\frac{826713271}{1188000000000} \\ -\frac{5094901033}{2376000000000} & 0 & -\frac{5094901033}{2376000000000} \\ -\frac{826713271}{1188000000000} & -\frac{5094901033}{2376000000000} & -\frac{826713271}{1188000000000} \end{bmatrix}.$$

The properties of this stencil can be found in the following

Theorem 5.1. *Let $A_{QSFEM}^{interior}$ be defined by (5.7). This stencil has the property that the functions $r_m(\beta)$ defined by (5.3,5.4) satisfy*

$$r_1(\beta) \equiv r_2(\beta) \equiv 0$$

$$r_3(\beta) = \frac{\cos(8\beta)}{774144}.$$

The quantity $\mathcal{D}\left(A_{QFEM}^{interior}\right)$ can be estimated by

$$\mathcal{D}\left(A_{QSFEM}^{interior}\right) \leq 1.3 \cdot 10^{-6} (kh)^7 + O\left((kh)^9\right).$$

Proof. The proof is purely technical but simple, therefore we give here only an outline of it.

First one has to expand the quotients $4\frac{\left(A_{QSFEM}^{interior}\right)_1}{\left(A_{QSFEM}^{interior}\right)_0}$ and $4\frac{\left(A_{QSFEM}^{interior}\right)_2}{\left(A_{QSFEM}^{interior}\right)_0}$ according to (5.1) to obtain the coefficients λ_m and ι_m. The proof then is completed by computing the functions $r_m(\beta)$ using the recursive formula (5.4). ■

6. Numerical Results

In this section we will present the results of an implementation of different GFEM to approximate the following problem

$$-\Delta u - k^2 u = 0 \text{ in } \Omega := (0,1) \times (0,1) \tag{6.1}$$

with boundary conditions

$$iku + \frac{\partial u}{\partial n} = g \text{ on } \Gamma := \partial\Omega. \tag{6.2}$$

The function g depends on the parameter θ and is given by

$$g(x) := \begin{cases} i(k-k_2)\, e^{ik_1 x_1} & \text{if } x \in \Gamma_1 := (0,1) \times (0,0), \\ i(k+k_1)\, e^{i(k_1+k_2 x_2)} & \text{if } x \in \Gamma_2 := (1,1) \times (0,1), \\ i(k+k_2)\, e^{i(k_1 x_1+k_2)} & \text{if } x \in \Gamma_3 := (0,1) \times (1,1), \\ i(k-k_1)\, e^{ik_2 x_2} & \text{if } x \in \Gamma_4 := (0,0) \times (0,1) \end{cases}$$

with $(k_1, k_2) = k(\cos\theta, \sin\theta)$.

The exact solution of this problem is

$$u_{ex}(x) := e^{i(k_1 x_1 + k_2 x_2)}.$$

6.1. Discretization techniques for the Helmholtz equation

We discretize the domain Ω by squares of side length $h = \frac{1}{n-1}$ and use bilinear elements. Consequently the system matrix has dimension n^2. We implemented the following three discretization methods.

1 Galerkin Finite Element Method

Writing (6.1) in a variational formulation and incorporating the boundary conditions results in the following problem.

Find $u \in \mathcal{H}^1(\Omega)$ such that

$$\int_\Omega \nabla u \nabla v - k^2 uv dx + ik \int_\Gamma uv d\Gamma_x = \int_\Gamma gv d\Gamma_x$$

is satisfied for all $v \in \mathcal{H}^1(\Omega)$. Replacing the infinite-dimensional space $\mathcal{H}^1(\Omega)$ by the finite element space $\mathcal{S}_h$ of bilinear elements and introducing the usual local basis results in a system of linear equations for the coefficients $\mathbf{u}$ of the basis representation (3.3). This method can also be described in terms of stencils. The interior stencil for the Galerkin method is given by

$$A_{gal}^{interior} := \begin{bmatrix} -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{8}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \end{bmatrix} - (kh)^2 \begin{bmatrix} \frac{1}{36} & \frac{1}{9} & \frac{1}{36} \\ \frac{1}{9} & \frac{4}{9} & \frac{1}{9} \\ \frac{1}{36} & \frac{1}{9} & \frac{1}{36} \end{bmatrix}.$$

A similar computation as explained in the proof of Theorem 5.1 yields the following estimate for the quantity $\mathcal{D}(stencil)$ which describes the approximation quality of this method for the Helmholtz equation.

$$\mathcal{D}\left(A_{gal}^{interior}\right) = (kh)^3 \max_{-\pi \le \beta < \pi} \frac{3 + \cos(4\beta)}{96} + O\left((kh)^5\right) = \frac{(kh)^3}{24} + O\left((kh)^5\right).$$

2 Generalized Least Squares Finite Element method (GLS-FEM)

In [9], Thompson and Pinsky have generalized the GLS-FEM, originally introduced by Harari and Hughes in [4] for a one-dimensional model problem, to the two space dimensions. Applied to problem (6.1) discretized by bilinear elements this method can be written in the form:

Find $u \in \mathcal{S}_h$ such that

$$\int_\Omega \nabla u \nabla v - \tau uv dx + ik \int_\Gamma uv d\Gamma_x = \int_\Gamma gv d\Gamma_x$$

is satisfied for all $v \in \mathcal{S}_h$. The parameter $\tau = \tau(k,h)$ is given by

$$\tau(k,h) = 6 \frac{4 - \cos\left(kh\cos\frac{\pi}{8}\right) - \cos\left(kh\sin\frac{\pi}{8}\right) - 2\cos\left(kh\cos\frac{\pi}{8}\right)\cos\left(kh\sin\frac{\pi}{8}\right)}{\left(2 + \cos\left(kh\cos\frac{\pi}{8}\right)\right)\left(2 + \cos\left(kh\sin\frac{\pi}{8}\right)\right)h^2}.$$

The interior stencil for the GLS-FEM can be written in the form

$$A_{GLS-FEM}^{interior} := \begin{bmatrix} -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{8}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \end{bmatrix} - h^2\tau(k,h) \begin{bmatrix} \frac{1}{36} & \frac{1}{9} & \frac{1}{36} \\ \frac{1}{9} & \frac{4}{9} & \frac{1}{9} \\ \frac{1}{36} & \frac{1}{9} & \frac{1}{36} \end{bmatrix}.$$

To compare this method with the standard Galerkin-FEM we state that

$$-h^2\tau = -(kh)^2 - \frac{1}{16}(hk)^4 + O\left((kh)^6\right)$$

which means that the GLS-FEM for bilinear elements is a higher order modification of the interior stencil of the Galerkin FEM. The approximation quality of this method to the Helmholtz equation in terms of $\mathcal{D}\,(stencil)$ is given by

$$\mathcal{D}\left(A^{interior}_{GLS-FEM}\right) = (kh)^3 \max_{-\pi\leq\beta<\pi}\left|\frac{\cos(4\beta)}{96}\right| + O\left((kh)^5\right) = \frac{(kh)^3}{96} + O\left((kh)^5\right)$$

and hence the pollution is essentially the same as for the Galerkin FEM.

3. Quasi-Stabilized Finite Element Method (QSFEM)

The interior stencil of this method was already presented in Section 5. Here we describe only the modeling of the boundary conditions and the assembling of the vector of the right-hand side. This is done analogously to the finite difference method by replacing the normal derivatives by a difference formula centered at the edge points and then eliminating the fictitious points. This technique is described together with an error analysis in [3, Section 4.7.2].

We recall that the approximation quality of this method was proved to be

$$\mathcal{D}\left(A^{interior}_{QSFEM}\right) = (kh)^7 \max_{-\pi\leq\beta<\pi}\left|\frac{\cos 8\beta}{774144}\right| + O\left((kh)^9\right) = \frac{\cos 8\beta}{774144} + O\left((kh)^9\right). \quad (6.3)$$

6.2. Numerical evaluation of the GFE methods for the Helmholtz equation

In order to measure the accuracy and compare the presented methods we have to introduce suitable norms. A natural measure of the approximation error of (6.1) is the energy norm, i.e. the usual $\mathcal{H}^1$-seminorm

$$e_1 := \frac{\|u_{ex} - u_{fe}\|_1}{\|u_{ex}\|_1}$$

with

$$\|u\|_1^2 := \int_\Omega \nabla u\,(x)\cdot\nabla\bar{u}\,(x)\,dx.$$

Alternatively we will measure the accuracy of the solution in the $\mathcal{L}^2$-norm

$$e_0 := \frac{\|u_{ex} - u_{fe}\|_0}{\|u_{ex}\|_0}$$

with

$$\|u\|_0^2 := \int_\Omega u\,(x)\,\bar{u}\,(x)\,dx.$$

Recalling the one-dimensional results (cf. Section 2) we expect that the error of the best approximation behaves like

$$\frac{\left\|u_{ex} - u_{opt}^1\right\|_1}{\|u_{ex}\|_1} \leq Chk,$$

and

$$\frac{\left\|u_{ex} - u_{opt}^0\right\|_0}{\|u_{ex}\|_0} \leq C\,(hk)^2\,.$$

In the one-dimensional case, the error of the Galerkin-FEM could be estimated by

$$e_1 \leq C\,(kh) + \tilde{C}k^2h\,(kh)$$

and

$$e_0 \leq C\,(hk)^2 + \tilde{C}k\,(kh)^2\,. \tag{6.4}$$

Obviously, the pollution of the $\mathcal{H}^1$-error becomes negligible if k^2h is small, but in the pre-asymptotic range we expect that the Galerkin solution differs substantially from the best approximation.

For the $\mathcal{L}^2$-error the pollution term is the dominant term in the error estimate for all values of h. We expect that, with increasing value of k, the distance of the graph of the Galerkin error from the best approximation error increases.

The error of the GFE solution depends on the direction θ of the wave vector $\mathbf{k}$, where

$$\mathbf{k} = (k_1, k_2)^T = k\,(\cos\theta, \sin\theta)\,.$$

It turns out that for special values of θ the GFE solutions effectively coincide with the best approximation, where for other values of θ the GFE solution differs substantially from the best approximation. Unfortunately, the direction θ is not known a priori. In order to guarantee that the error of the GFE solution is under control (i.e. the solution is reliable) one has to assume that the error is sufficiently small even if the solution would correspond to the value of θ with the largest error. Therefore for $j \in \{0, 1\}$, we have computed the quantity

$$e_j^{\max} = \max_{-\pi \leq \omega < \pi} e_j\,(\omega)\,.$$

Here, $e_j\,(\omega)$ denotes the error of the GFE solution of problem (6.1,6.2), where the parameter θ for the right-hand side g is chosen as ω. The exact solution in this

case is given by

$$u_{ex}(x) = e^{ik(x_1 \cos\omega + x_2 \sin\omega)}.$$

We approximate the maximum above by choosing the set of ω-values

$$\Xi = \left\{0, \frac{\pi}{16}, \frac{\pi}{8}, \frac{3\pi}{16}, \frac{\pi}{4}, \frac{3\pi}{8}\right\}$$

Due to the rotational symmetry of our problem we know that the error corresponding to a direction θ is the same as for $\theta + m\frac{\pi}{2}$ for integers m. The maximum above is approximated by

$$e_j^{\max} \approx \max_{\omega \in \Xi} e_j(\omega).$$

We restrict the step size h to

$$hk \leq \frac{\pi}{2}.$$

This assumption is natural because it guarantees at least four elements per wave length (see Fig. 6.1).

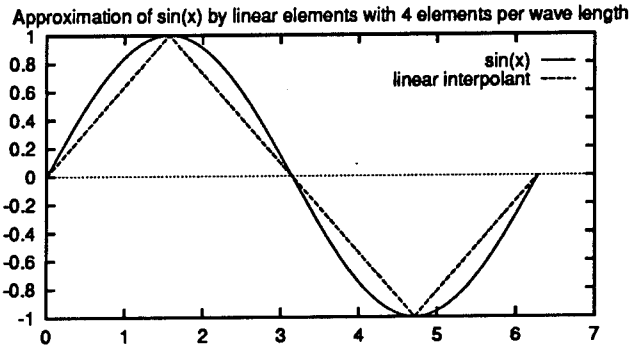


Figure 6.1: Approximation of sin(x) by the piecewise linear interpolant with four elements per wave length.

For larger values of h the error of the best approximation would be too large for practical applications.

Figure 6.2 and 6.5 show the $\mathcal{H}^1$-error, resp. $\mathcal{L}^2$-error of the three discretization schemes and of the best approximation for $k = 30, 100, 150$. The plots consists of groups of four lines, where the lowest line always corresponds to the best approximation, the next one corresponds to the QSFEM, the third belongs to the

GLS-FEM and the highest line always corresponds to the Galerkin solution. We see that the error of the Galerkin solution behaves as expected. The $\mathcal{H}^1$-error differs substantially from the best approximation in the pre-asymptotic range, while this range increases for larger values of k. On the other hand the pollution becomes negligible if h is sufficiently small, since. k^2h is small. The situation for the $\mathcal{L}^2$-norm is different. Here, the pollution does not vanish as $h \to 0$. The lines becomes parallel (in our log-log plot) for small values of h whereas the distance from the best approximation increases with higher wave number k for all h in accordance with the theoretical estimate (6.4).

The improvement of the QSFEM is obvious. The size and range of the pollution even for $k = 150$ is very small and remains nearly constant for different values of k. For the $\mathcal{L}^2$-norm this behavior can be observed from the constant (w.r.t. the wave number k) distance between the graph of the best approximation and the QSFE-solution.

The GLS-FEM shows nearly no improvement over the Galerkin FEM for relatively large values of $hk \sim \pi/2$, while the pollution decreases faster with respect to h than for the Galerkin FEM. These numerical results are in accordance with the different sizes of the theoretical quality measure $\mathcal{D}$ (*stencil*) computed above. The pictures (6.2-6.5) can also be interpreted as follows. Let us assume that we want to approximate the solution of our model problem with an relative error of ϵ in the $\mathcal{L}^2$-norm. Then we can ask how many degrees of freedom (DOF) are necessary to get this accuracy. The following table shows this dependency for some values of ϵ.

DOF necessary to obtain an accuracy of ϵ in the $\mathcal{L}^2$-norm								
	$k = 100$				$k = 150$			
ϵ	BA	QSFEM	GLSFEM	FEM	BA	QSFEM	GLSFEM	FEM
30%	45^2	63^2	142^2	279^2	68^2	97^2	258^2	512^2
10%	71^2	100^2	248^2	485^2	105^2	148^2	449^2	879^2
5%	92^2	140^2	357^2	685^2	140^2	209^2	634^2	1240^2

As mentioned above the error of the GFE-solution depends significantly on the wave direction θ. In the following plots we illustrate the pollution effect dependent of the parameter θ. For constant kh we know that the error of the best approximation in the $\mathcal{H}^1$-norm is of order kh. We have chosen $kh = 1.5, 0.7, 0.3$, where $kh = 1.5 \approx \pi/2$ corresponds to four elements per wave length, $kh = 0.7$ to eight and $kh = 0.3$ to 16 elements per wave length. In Figures 6.6-6.14, we

have plotted the difference of the GFE-errors from the best approximation error. We see that the Galerkin FEM is relatively independent of the parameter θ but the difference from the best approximation is significantly for all considered cases. The difference of the GLS-FE solution from the best approximation is practically zero for $\theta = \frac{\pi}{8}, \frac{3\pi}{8}$, where for $\theta = 0, \frac{\pi}{4}, \frac{\pi}{2}$ it is especially for $kh = \pi/2$ practically as bad as the Galerkin FEM. The scaling of the axes of the GLS-FEM plots and the QSFEM plots is always the same. We see that for the considered range of k the QSFEM has practically no pollution. The difference from the best approximation is small for all considered values of θ and is fairly steady with the wave number k. In contrast to, e.g. $\theta = 0$, the difference of the GLS-FE-solution from the best approximation increases with higher wave number k which means for this method it is not sufficient to restrict the quantity kh to get a small error. Figures 6.15-6.23 shows the dependency of the pollution on kh and θ in the $\mathcal{L}^2$-norm. One can observe that the pollution of the QSFEM is even smaller than in the $\mathcal{H}^1$-norm, where the behavior of the GLS-FEM and Galerkin FEM is quite similar.

6.3. Conclusions and Remarks

In this subsection we summarize and comment on the numerical results and relate them to the theory presented in the previous sections. In Section 2 we had analyzed a one-dimensional example which shows the typical pollution effect for the GFE-approximations. We have seen that if we are able to design a GFEM which has no pollution for this example it will also have no pollution for more general cases (cf. Theorem 2.7) if we treat the boundary conditions and an inhomogenous right-hand side in an appropriate way. On the other hand we have proven that if the discrete wave number and the exact one are not the same, then the pollution cannot be avoided by any treatment of the boundary conditions.

In the two-dimensional case we have seen that a necessary requirement to get a small pollution is to minimize the maximal distance $\mathcal{D}$ of the zeros of the discrete symbol and of the operator symbol. We were able to design a GFEM called QSFEM such that $\mathcal{D}$ is minimal. On the other hand we have not investigated a special treatment of the boundary conditions. Nevertheless, from the numerical results presented above it follows that even a straightforward treatment of the boundary condition (centered differences) results in a GFEM with a negligible pollution effect for moderate wave numbers. We conclude that on the one hand the pollution effect for the FE treatment of the Helmholtz equation is not avoidable

in principle, while on the other hand it is possible to design the QSFEM having a pollution which is not visible in practice.

Acknowledgments. This work was supported by ONR Grant N00014-93-I-0131. The second author was supported by Grant No 517 402 524 3 of the German Academic Exchange Service (DAAD). The work of the fourth author was supported by the German Research Foundation (DFG), Grant No.Sa 607/1-1.

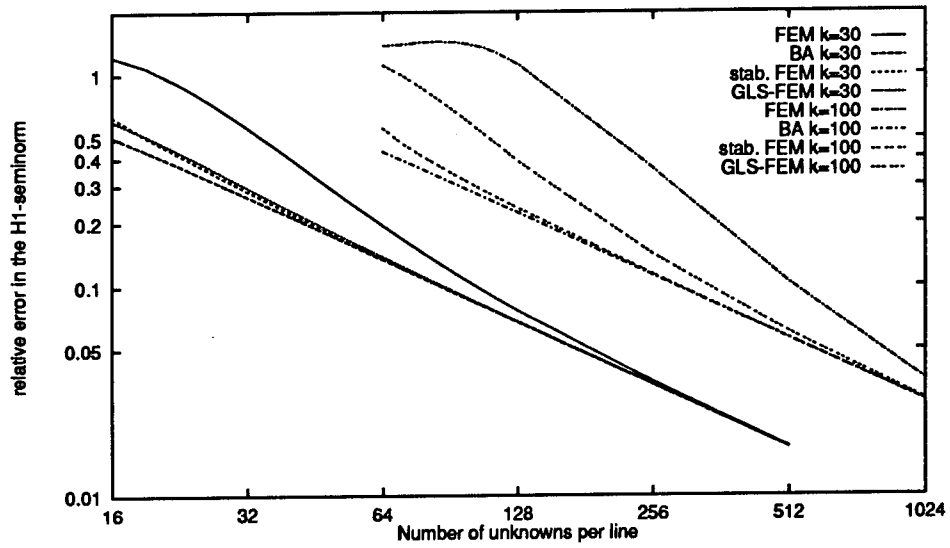


Figure 6.2: Relative error in the H^1-seminorm for $k = 30, 100$

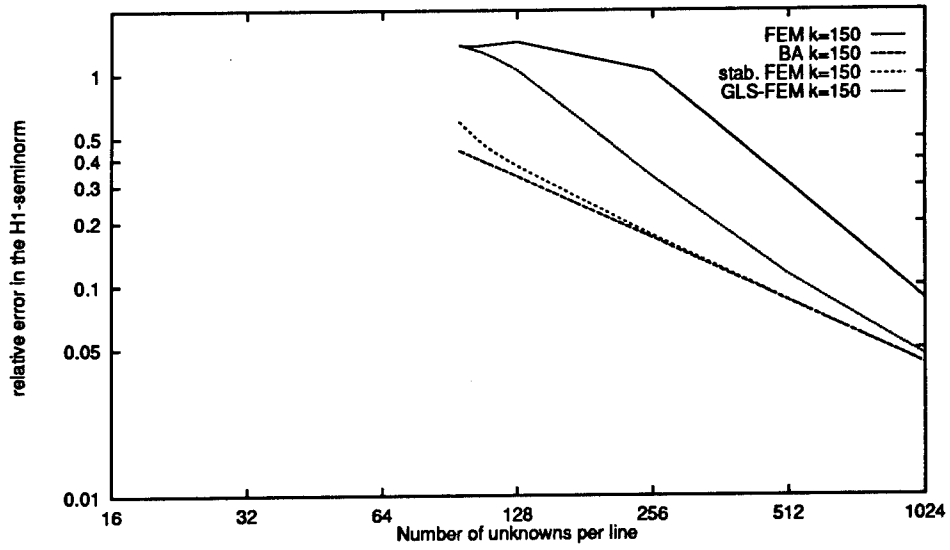


Figure 6.3: Relative error in H^1-seminorm for $k = 150$

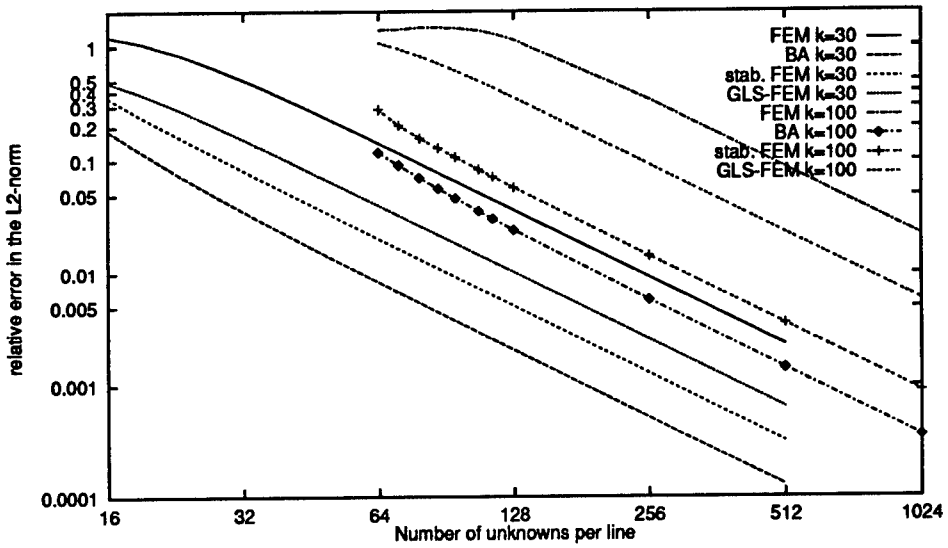


Figure 6.4: Relative error in the L^2-norm for $k = 30, 100$

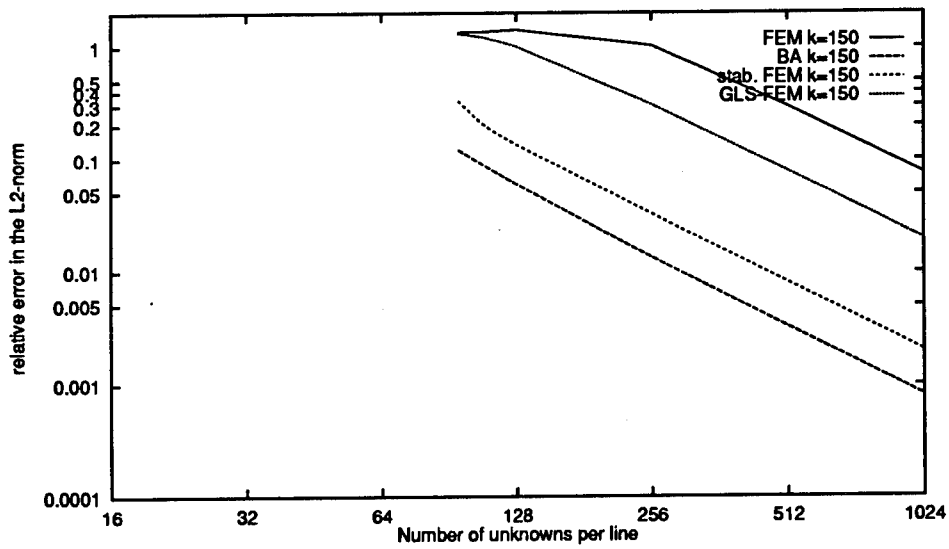


Figure 6.5: Relative error in the L^2-norm for $k = 150$

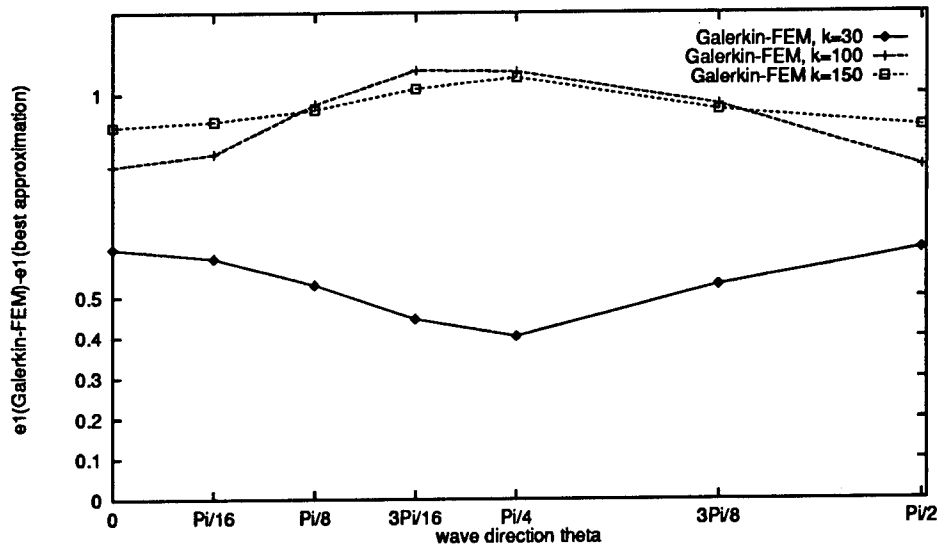


Figure 6.6: Dependency of the H^1-error on the angle θ for $kh = 1.5$ for the Galerkin-FEM

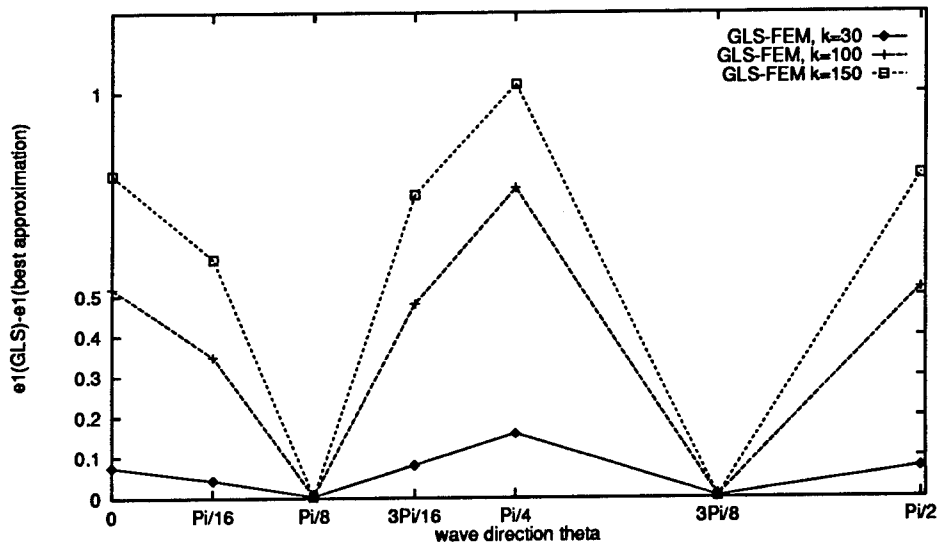


Figure 6.7: Dependency of the H^1-error on the angle θ for $kh = 1.5$ for the GLS-FEM

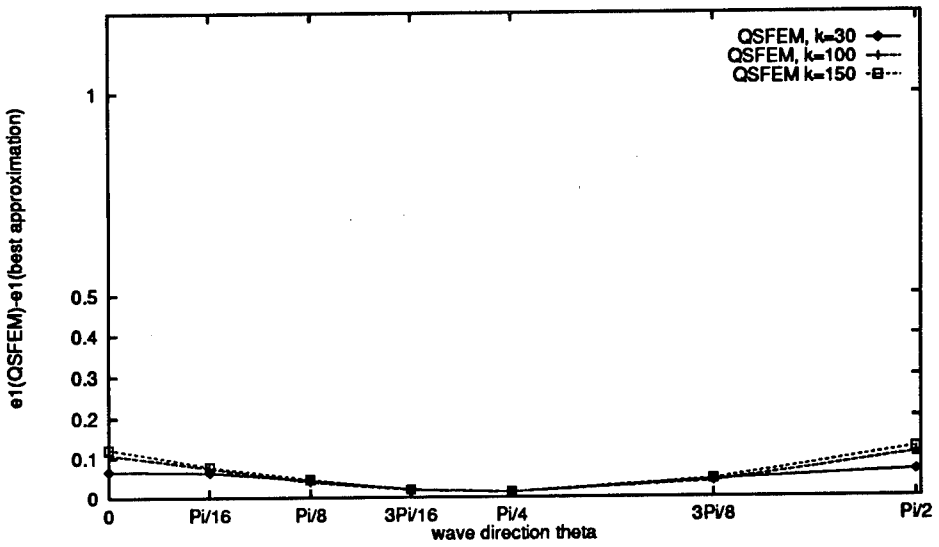


Figure 6.8: Dependency of the H^1-error on the angle θ for $kh = 1.5$ for the QSFEM

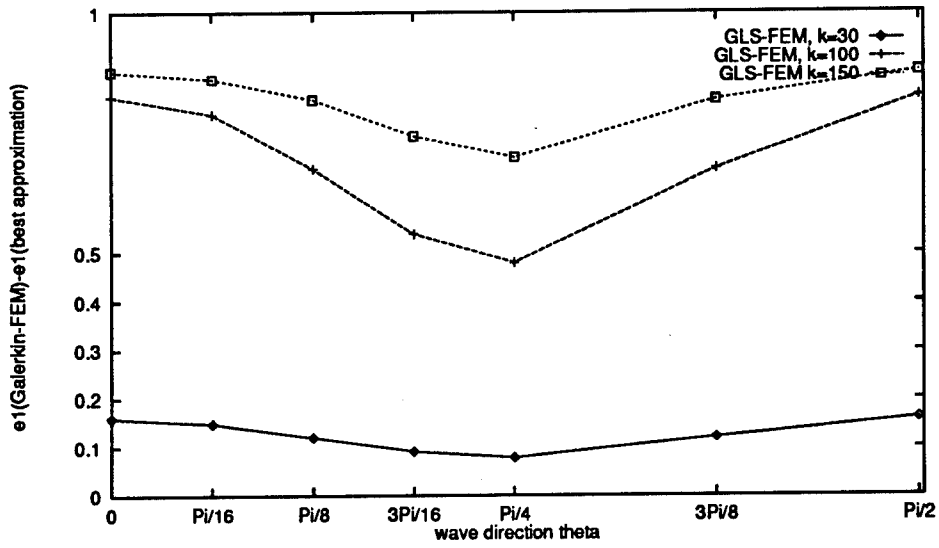


Figure 6.9: Dependency of the H^1-error on the angle θ for $kh = 0.7$ for the Galerkin-FEM

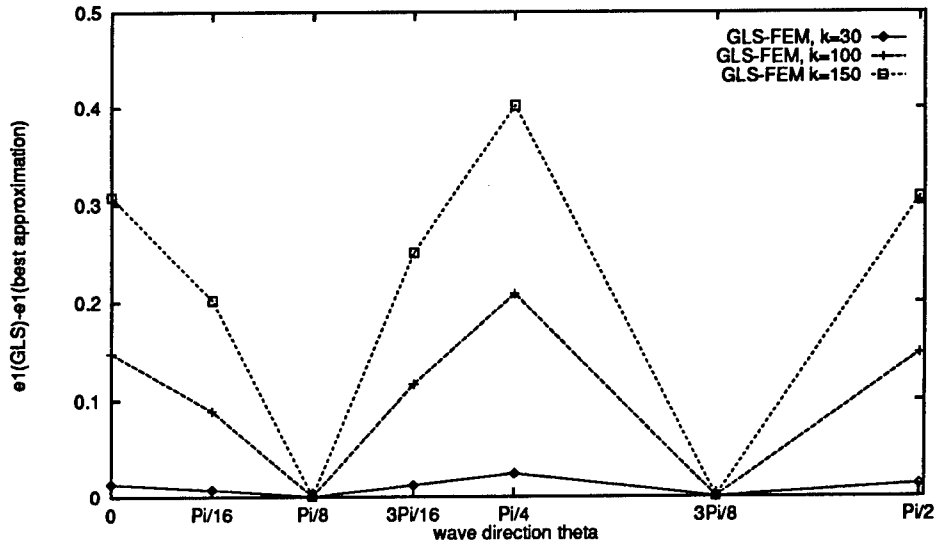


Figure 6.10: Dependency of the H^1-error on the angle θ for $kh = 0.7$ for the GLS-FEM

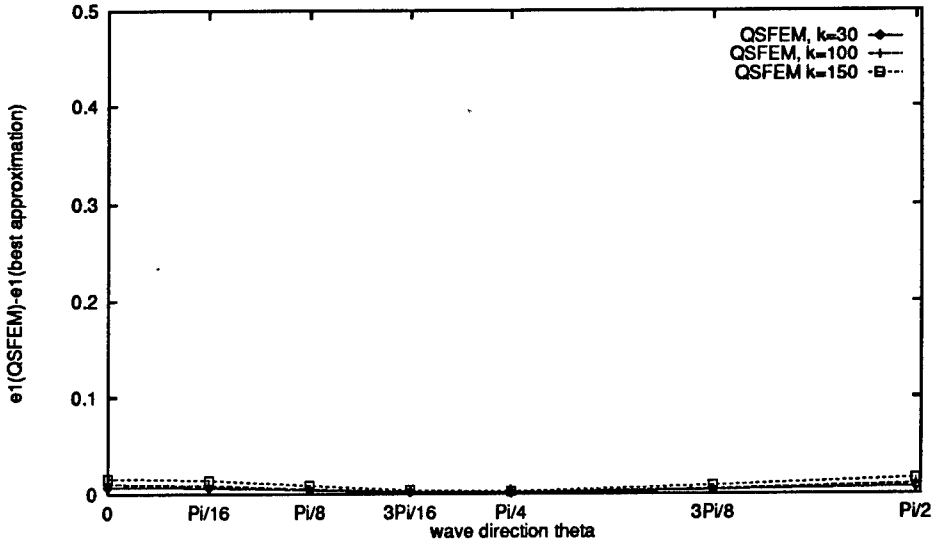


Figure 6.11: Dependency of the H^1-error on the angle θ for $kh = 0.7$ for the QSFEM

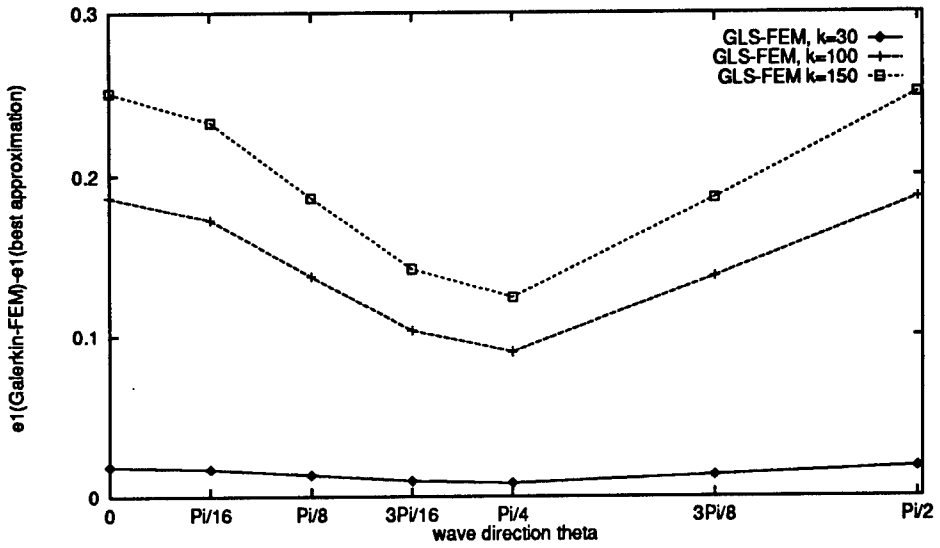


Figure 6.12: Dependency of the H^1-error on the angle θ for $kh = 0.3$ for the Galerkin-FEM

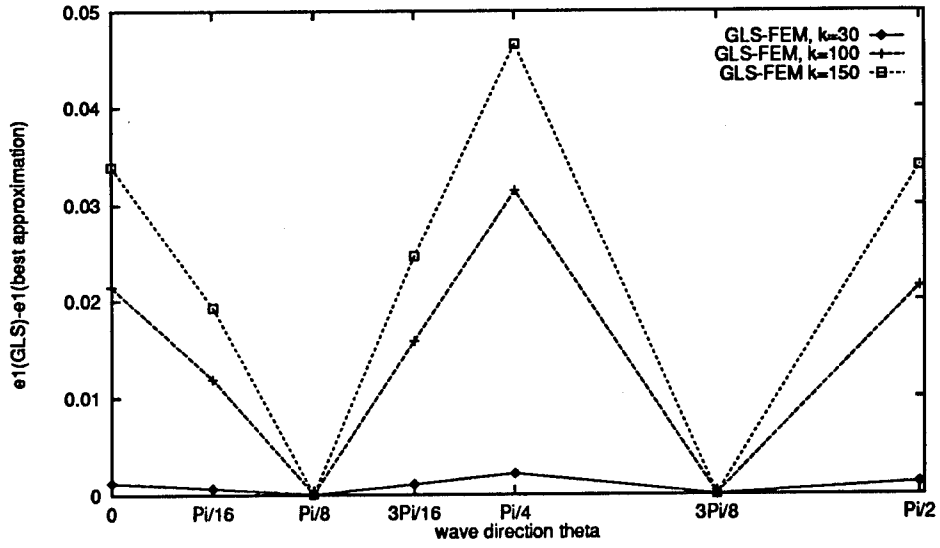


Figure 6.13: Dependency of the H^1-error on the angle θ for $kh = 0.3$ for the GLS-FEM

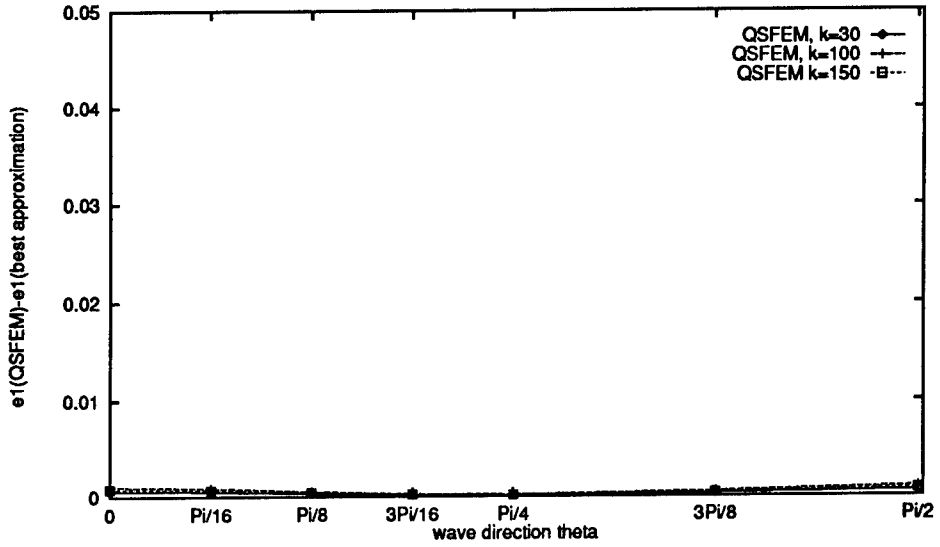


Figure 6.14: Dependency of the H^1-error on the angle θ for $kh = 0.3$ for the QSFEM

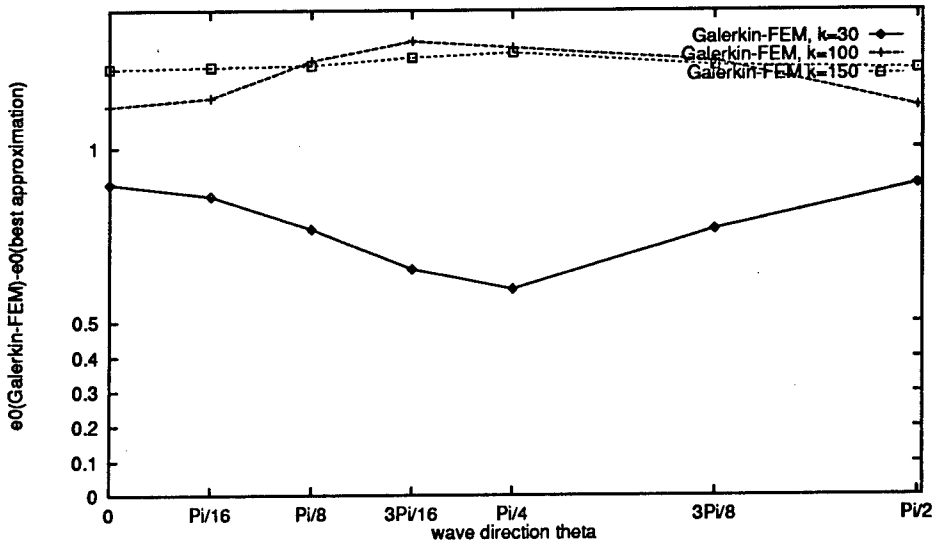


Figure 6.15: Dependency of the L^2-error on the angle θ for $kh = 1.5$ for the Galerkin-FEM

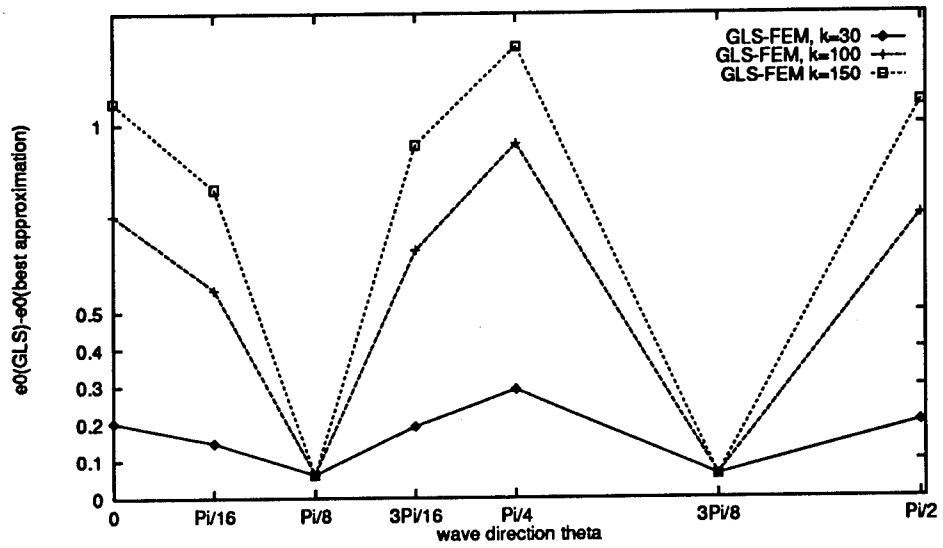


Figure 6.16: Dependency of the L^2-error on the angle θ for $kh = 1.5$ for the GLS-FEM

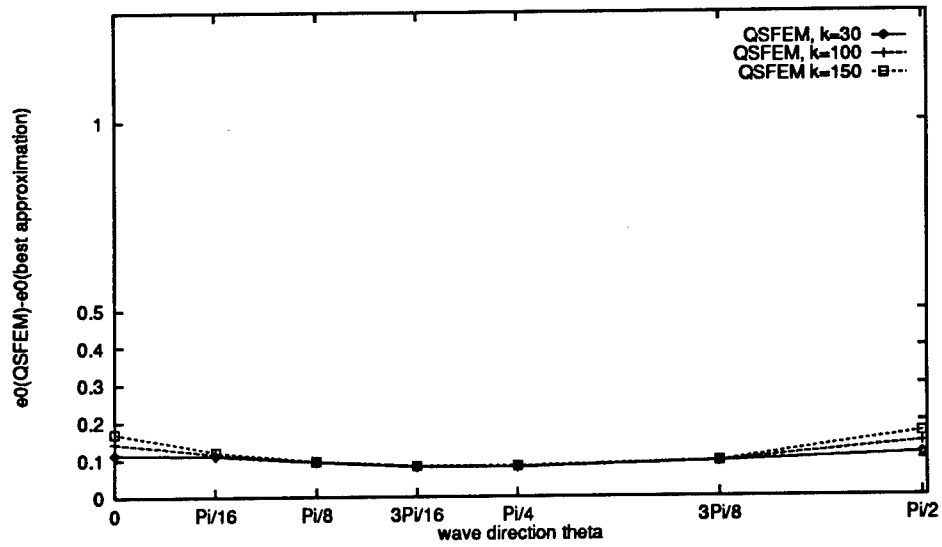


Figure 6.17: Dependency of the L^2-error on the angle θ for $kh = 1.5$ for the QSFEM

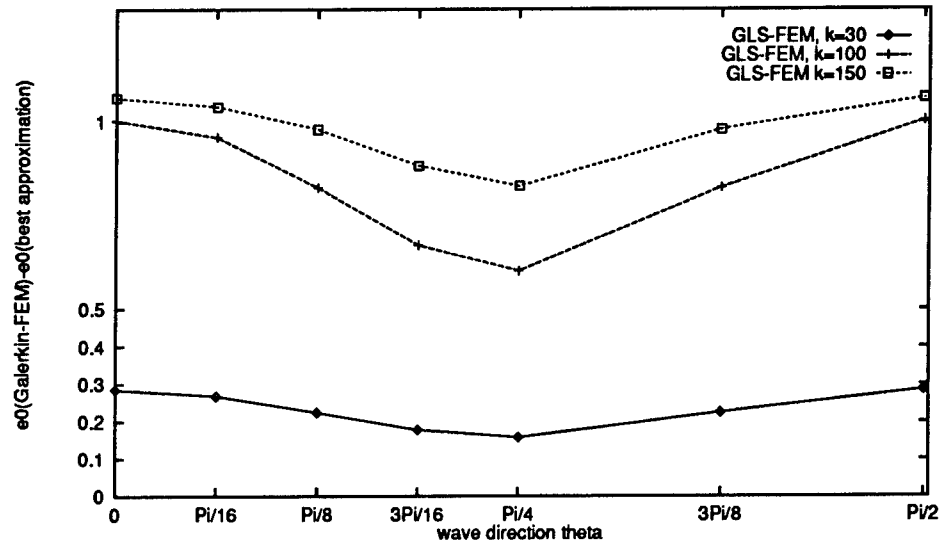


Figure 6.18: Dependency of the L^2-error on the angle θ for $kh = 0.7$ for the Galerkin-FEM

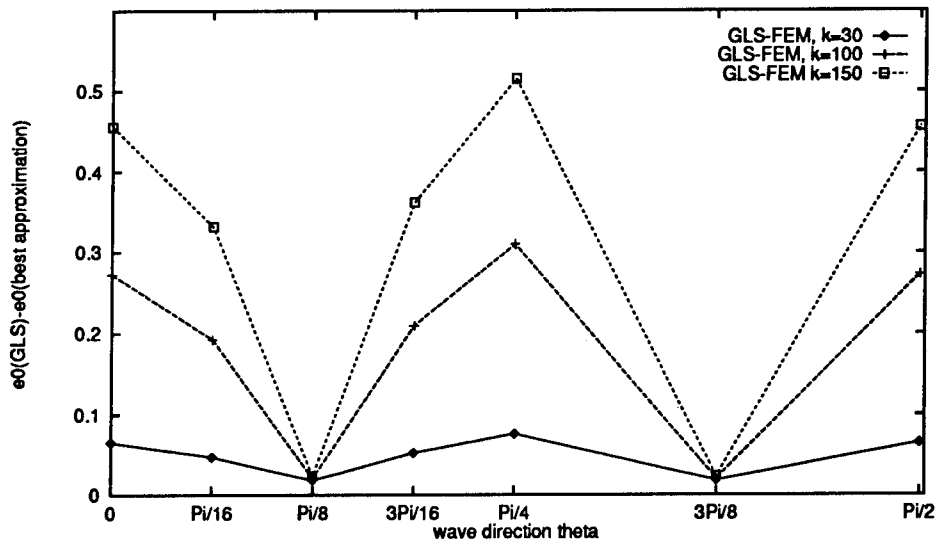


Figure 6.19: Dependency of the L^2-error on the angle θ for $kh = 0.7$ for the GLS-FEM

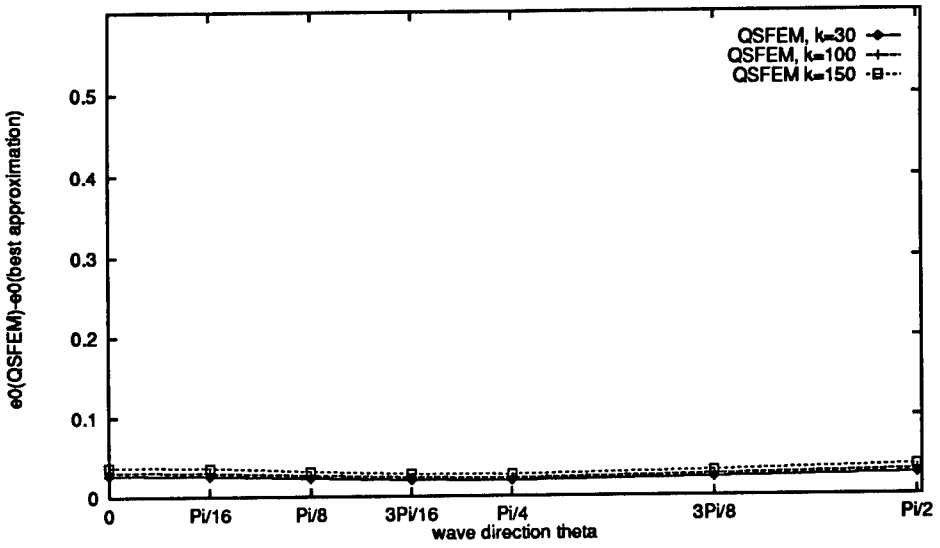


Figure 6.20: Dependency of the L^2-error on the angle θ for $kh = 0.7$ for the QSFEM

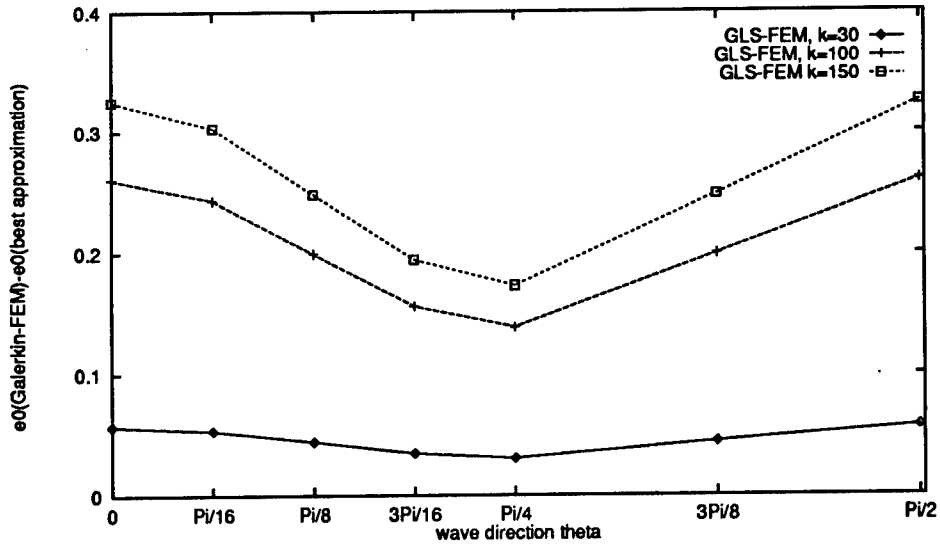


Figure 6.21: Dependency of the L^2-error on the angle θ for $kh = 0.3$ for the Galerkin-FEM

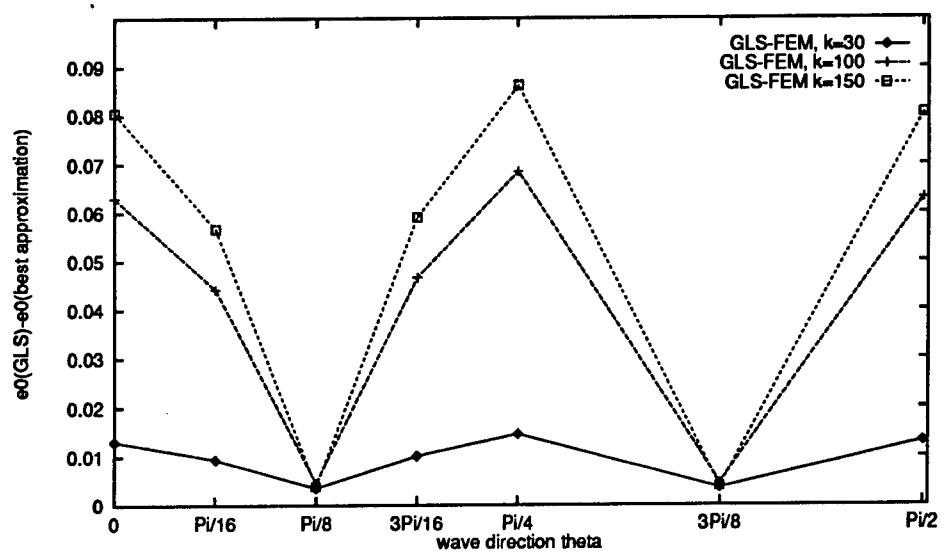


Figure 6.22: Dependency of the L^2-error on the angle θ for $kh = 0.3$ for the GLS-FEM

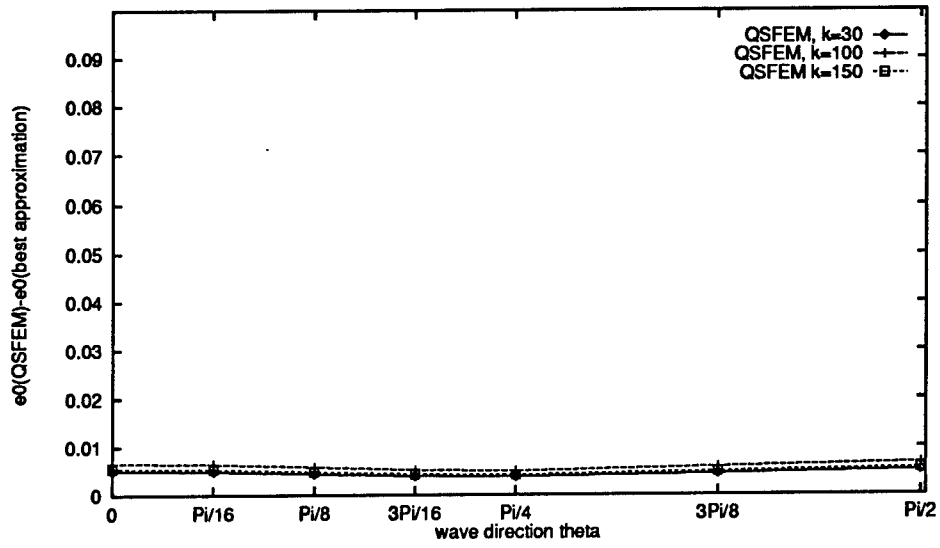


Figure 6.23: Dependency of the L^2-error on the angle θ for $kh = 0.3$ for the QSFEM

References

[1] I. Babuška and J.E. Osborn. Generalized Finite Element Methods: Their performance and their relation to mixed methods. *SIAM, J. Numer. Anal.*, 20(3):510–536, 1983.

[2] I.M. Babuška and S.A. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers. Technical Report BN-1172, IPST, University of Maryland at College Park, 1994.

[3] W. Hackbusch. *Elliptic Differential Equations.* Springer Verlag, 1992.

[4] I. Harari and T.J.R. Hughes. Finite element methods for the Helmholtz equation in an exterior domain: model problems. *Computer methods in applied mechanics and engineering*, 87:59–96, North Holland, 1991.

[5] F. Ihlenburg and I. Babuška. Finite element solution to the Helmholtz equation with high wave number. Part I: The h-version of the FEM. Technical Report BN-1159, IPST, University of Maryland at College Park, 1993.

[6] F. Ihlenburg and I. Babuška. Dispersion analysis and error estimation of Galerkin finite element methods for the numerical computation of waves. Technical Report BN-1174, IPST, University of Maryland at College Park, 1994.

[7] F. Ihlenburg and I. Babuška. Finite element solution to the Helmholtz equation with high wave number. Part II: The h-p version of the FEM. Technical Report BN-1173, IPST, University of Maryland at College Park, 1994.

[8] J.B. Keller and D. Givoli. Exact non-reflecting boundary conditions. *J. Comp. Phys.*, 82:172–192, 1989.

[9] L.L.Thompson and P.M. Pinsky. A Galerkin least squares finite element method for the two-dimensionsal Helmholtz equation. *International Journal for Numerical Methods in Engineering*, to appear.

The Laboratory for Numerical Analysis is an integral part of the Institute for Physical Science and Technology of the University of Maryland, under the general administration of the Director, Institute for Physical Science and Technology. It has the following goals:

To conduct research in the mathematical theory and computational implementation of numerical analysis and related topics, with emphasis on the numerical treatment of linear and nonlinear differential equations and problems in linear and nonlinear algebra.

To help bridge gaps between computational directions in engineering, physics, etc., and those in the mathematical community.

To provide a limited consulting service in all areas of numerical mathematics to the University as a whole, and also to government agencies and industries in the State of Maryland and the Washington Metropolitan area.

To assist with the education of numerical analysts, especially at the postdoctoral level, in conjunction with the Interdisciplinary Applied Mathematics Program and the programs of the Mathematics and Computer Science Departments. This includes active collaboration with government agencies such as the National Institute of Standards and Technology.

To be an international center of study and research for foreign students in numerical mathematics who are supported by foreign governments or exchange agencies (Fulbright, etc.).

Further information may be obtained from **Professor I. Babuška**, Chairman, Laboratory for Numerical Analysis, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742-2431.